Computer Science Department

TECHNICAL REPORT

Computing Stable Eigendecompositions of Matrix Pencils

James Weldon Demmel
Courant Institute of Mathematical Sciences
251 Mercer Str.
New York, NY 10012

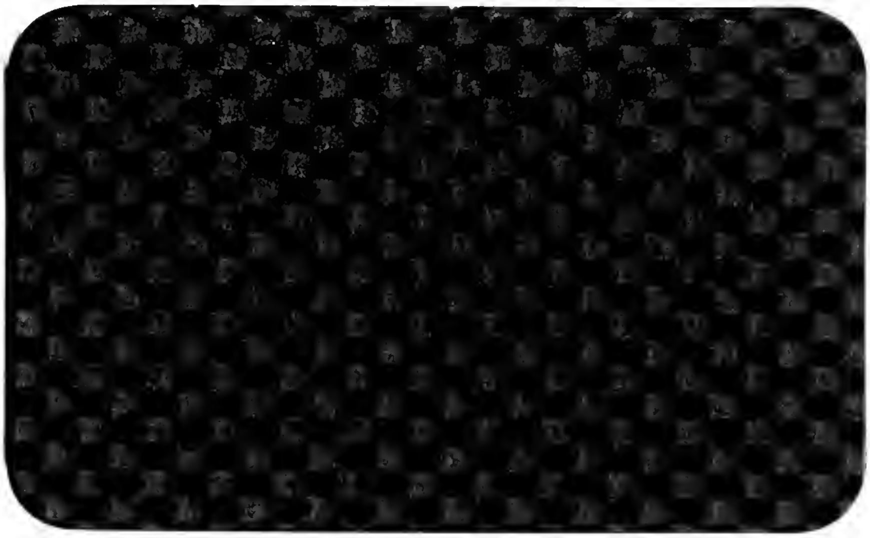
Bo Kågström
Institute of Information Processing
University of Umeå
S-901 87 Umeå
Sweden

May 1985

Technical Report 163

NEW YORK UNIVERSITY

Department of Computer Science
Courant Institute of Mathematical Sciences
251 MERCER STREET, NEW YORK, N.Y. 10012



Computing Stable Eigendecompositions of Matrix Pencils

James Weldon Demmel
Courant Institute of Mathematical Sciences
251 Mercer Str.
New York, NY 10012

Bo Kågström
Institute of Information Processing
University of Umeå
S-901 87 Umeå
Sweden

May 1985

Technical Report #163

Abstract

If a matrix pencil $A-\lambda B$ is known only to within a tolerance ϵ (because of measurement or round-off errors), then it may be difficult to compute a generalized eigendecomposition of $A-\lambda B$ since its eigenspaces are discontinuous functions of its entries. We are interested in computing an eigendecomposition of $A-\lambda B$ which varies continuously and boundedly as $A-\lambda B$ varies inside a ball of radius ϵ. There are two cases with qualitatively different solutions. The first case is when $A-\lambda B$ is regular, i.e. $\det(A-\lambda B)$ is not identically zero. In this case we show how to partition the spectrum of $A-\lambda B$ into disjoint pieces which remain disjoint and whose associated eigenspaces vary smoothly. The second case is when $A-\lambda B$ is singular (i.e. either $\det(A-\lambda B)=0$ or $A-\lambda B$ is nonsquare). This case is more difficult than the first because applications call for computing nongeneric eigenspaces which exist only when $A-\lambda B$ lies in a proper variety (a set of measure zero). The known algorithms for computing these nongeneric structures produce the eigendecomposition of a pencil close to the input where this perturbed pencil is guaranteed to lie in the proper variety mentioned above. In this case we prove that as long as the norm of the perturbations produced by the algorithm are smaller than a certain ϵ we can compute from the pencil, the resulting nongeneric eigenspaces produced by the algorithm vary smoothly. We illustrate this theorem with an example from systems theory: we derive perturbation bounds for the controllable subspace and uncontrollable modes of a system $\dot{x}=Cx+Du$.

1. Introduction.

If we are given a complex m by n matrix pencil $A-\lambda B$ which we only know to within a tolerance $\epsilon>0$, what does it mean to compute an eigendecomposition of $A-\lambda B$? By only knowing $A-\lambda B$ to a tolerance ϵ we mean that $A-\lambda B$ is indistinguishable from any pencil in the set

$$\mathrm{P}(\epsilon) \equiv \{A+E-\lambda(B+F):\|(E,F)\|_E < \epsilon\} \ .$$

An eigendecomposition of $A-\lambda B$ will be written

$$A-\lambda B = P(S-\lambda T)Q^{-1} \tag{1.1}$$

where P is an m by m nonsingular matrix, Q is an n by n nonsingular matrix, and S and T are block diagonal: $S=\mathrm{diag}(S_{11}, \ldots, S_{bb})$ and $T=\mathrm{diag}(T_{11}, \ldots, T_{bb})$. We can group the columns of P into blocks corresponding to the blocks of $S-\lambda T$: $P=[P_1 | \cdots | P_b]$ where P_i is m by m_i, m_i being the number of rows of $S_{ii}-\lambda T_{ii}$. Similarly, we can group the columns of Q into blocks corresponding to the blocks of $S-\lambda T$: $Q=[Q_1| \cdots |Q_b]$ where Q_i is n by n_i, n_i being the number of columns of $S_{ii}-\lambda T_{ii}$.

The diagonal blocks $S_{ii}-\lambda T_{ii}$ contain information about the generalized eigenstructure of the pencil $A-\lambda B$ and P_i and Q_i contain information about the corresponding generalized eigenspaces. One canonical decomposition of the form (1.1) we will refer to is the *Kronecker Canonical Form* or KCF [Gantmacher], where each block $S_{ii}-\lambda T_{ii}$ must be of one of the following forms:

$$S_{ii}-\lambda T_{ii} = \begin{bmatrix} \lambda_0-\lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_0-\lambda \end{bmatrix}$$

This is simply a Jordan block. λ_0 is called a finite eigenvalue of $A-\lambda B$.

$$S_{ii}-\lambda T_{ii} = \begin{bmatrix} 1 & \lambda & & \\ & \ddots & \ddots & \\ & & \ddots & \lambda \\ & & & 1 \end{bmatrix}$$

This block corresponds to an infinite eigenvalue of multiplicity equal to the dimension of the block. The blocks of finite and infinite eigenvalues together constitute the *regular* part of the pencil.

$$S_{ii}-\lambda T_{ii} = L_k \equiv \begin{bmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \lambda & 1 \end{bmatrix}$$

This k by $k+1$ block is called a singular block of minimal right (or column) index k. It has a one dimensional right null space for any λ.

$$S_{ii}-\lambda T_{ii} = L_j^T \equiv \begin{bmatrix} \lambda & & \\ 1 & \ddots & \\ & \ddots & \lambda \\ & & 1 \end{bmatrix}$$

This $j+1$ by j block is called a singular block of minimal left (or row) index j. It has a one dimensional left null space for any λ. The left and right singular blocks together constitute the *singular* part of the pencil.

We would like to produce an eigendecomposition which is valid in some way for all pencils in P(ϵ), and gives as much information about all pencils in P(ϵ) as possible. The kind of information such a decomposition can provide will depend on whether the pencil $A-\lambda B$ is regular or singular. $A-\lambda B$ is *regular* if $A-\lambda B$ is square and $\det(A-\lambda B)\neq 0$ for some λ. This is equivalent to $A-\lambda B$ having only a regular part in its KCF. $A-\lambda B$ is *singular* if either $A-\lambda B$ is square and $\det(A-\lambda B)=0$ for all λ, or else $A-\lambda B$ is nonsquare. This is equivalent to $A-\lambda B$ having a singular part in its KCF [Gantmacher].

In the regular case, $A-\lambda B$ has n generalized eigenvalues which may be finite or infinite. The diagonal blocks of $S-\lambda T$ partition the spectrum of $A-\lambda B$ as follows:

$$\sigma=\sigma(A-\lambda B) = \bigcup_{l=1}^{b} \sigma(S_{ll}-\lambda T_{ll}) = \bigcup_{l=1}^{b} \sigma_l \ .$$

The subspaces spanned by P_l and Q_l are called *left and right deflating subspaces* of $A-\lambda B$ corresponding to the part of the spectrum σ_l [Stewart3,Van Dooren3]. As shown in [Van Dooren3], a pair of subspaces **P** and **Q** is deflating for $A-\lambda B$ if $\mathbf{P}=A\mathbf{Q}+B\mathbf{Q}$ and $\dim(\mathbf{Q})=\dim(\mathbf{P})$. They are the generalization of invariant subspaces for the standard eigenvalue problem $A-\lambda I$ to the regular pencil case: **Q** is a (right) invariant subspace of A if $\mathbf{Q}=A\mathbf{Q}+\mathbf{Q}$, i.e. $A\mathbf{Q}\subseteq\mathbf{Q}$.

To illustrate and motivate our approach to regular pencils, we indicate how we would decompose P(ϵ) for various values of ϵ, where $A-\lambda B$ is given by

$$A-\lambda B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \eta \end{bmatrix} - \lambda \begin{bmatrix} \eta & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

where η is a small number. It is easy to see that $\sigma=\{1/\eta, 0, \eta\}$. The three deflating subspaces corresponding to these eigenvalues are spanned by the three columns of the 3 by 3 identity matrix. For ϵ sufficiently small, the spectrum of any pencil in P(ϵ) will contain 3 points, one each inside disjoint sets centered at $1/\eta$, 0 and η. In fact, we can draw 3 disjoint closed curves surrounding 3 disjoint regions, one around each $\lambda\in\sigma$ such that each pencil in P(ϵ) has exactly one eigenvalue in the region surrounded by each closed curve. Similarly, the three deflating subspaces corresponding to each eigenvalue remain close to orthogonal. Thus, for ϵ sufficiently small, we partition σ into three sets, $\sigma_1=\{1/\eta\}$, $\sigma_2=\{0\}$ and $\sigma_3=\{\eta\}$.

As ϵ increases to $\eta/\sqrt{2}$, it becomes impossible to draw three such curves because there is a pencil almost within distance $\eta/\sqrt{2}$ of $A-\lambda B$ with a double eigenvalue at $\eta/2$:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \eta/2 & \zeta \\ 0 & 0 & \eta/2 \end{bmatrix} - \lambda \begin{bmatrix} \eta & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} ,$$

where ζ is an arbitrarily small nonzero quantity. Furthermore, there are no longer three independent deflating subspaces, because the ζ causes the two deflating subspaces originally belonging to 0 and η to merge into a single two-dimensional deflating subspace. We can, however, draw two disjoint closed curves, one around $1/\eta$ and the other around 0 and η, such that every pencil in P(ϵ) has one eigenvalue inside the curve around $1/\eta$ and two eigenvalues inside the other curve. In this case we partition $\sigma=\{0,\eta\} \bigcup \{1/\eta\} = \sigma_1 \bigcup \sigma_2$.

As ϵ increases to 1, it no longer becomes possible to draw two disjoint closed curves any more, but merely one around all three eigenvalues, since it is possible to find a pencil inside P(ϵ) with η and $1/\eta$ having merged into a single eigenvalue near 1, as well as another pencil inside P(ϵ) where 0 and η have merged into a single eigenvalue near $\eta/2$. In this case we cannot partition σ into any smaller sets.

This example motivates the definition of a *stable decomposition of a regular pencil*: the decomposition in (1.1) is stable if the entries of P, Q, S and T are continuous and bounded functions of the entries of A and B as $A-\lambda B$ varies inside P(ϵ). In particular, we insist the dimensions n_i of the $S_{ii}-\lambda T_{ii}$ remain constant for $A-\lambda B$ in P(ϵ). This corresponds to partitioning $\sigma=\bigcup_{i=1}^{b}\sigma_i$ into disjoint pieces which remain disjoint for $A-\lambda B$ in P(ϵ). We illustrated this disjointness in the example by surround each σ_i by its own disjoint closed curve. For numerical reasons we will also insist that the matrices P and Q in (1.1) have their condition numbers bounded by some (user specified) threshold TOL for all pencils in P(ϵ). This is equivalent to insisting that the deflating subspaces belonging to different σ_i not contain vectors pointing in nearly parallel directions.

In the above example, as ϵ grows to $\eta/\sqrt{2}$, there are pencils in P(ϵ) where the deflating subspaces belonging to η and 0 become nearly parallel:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \eta/2+\zeta^2 & \zeta \\ 0 & 0 & \eta/2 \end{bmatrix} - \lambda \begin{bmatrix} \eta & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} .$$

The two right deflating subspaces in question are spanned by $[0,1,0]^T$ and $[0,-1/\zeta,1]^T$, respectively, which become nearly parallel as ζ approaches 0. The numerical reason for constraining the condition numbers of P and Q is that they indicate approximately how much accuracy we expect to lose in computing the decomposition (1.1) [Demmel1]. Therefore the user might wish to specify a maximum condition number TOL he is willing to tolerate in a stable decomposition as well as specifying the uncertainty ϵ in his data.

With this introduction, we can explain our first main result which is a criterion for deciding whether a decomposition (1.1) is stable or not:

Theorem A: Let $A-\lambda B$, ϵ and TOL be given. Let $\sigma = \bigcup_{i=1}^{b}\sigma_i$ be some partitioning of σ into disjoint sets. Define x_i for $1 \le i \le b$ as

$$x_i \equiv \epsilon \, \frac{(2\cdot(p_i^2 + q_i^2))^{1/2} + 2\cdot\max\,(p_i,q_i)}{\min\,(\mathrm{Dif}_u(\sigma_i,\sigma-\sigma_i)\ ,\ \mathrm{Dif}_l(\sigma_i,\sigma-\sigma_i)} \quad , \tag{1.2}$$

where p_i, q_i, Dif_u and Dif_l will be explained below.

The corresponding decomposition (1.1) is stable if the following two criteria are satisfied:

$$\max_{1\le i\le b} x_i < 1 \tag{1.3}$$

and

$$2\cdot b\cdot \max_{1\le i,j\le b}\,(p_i\ ,\ q_j)\cdot \max_{1\le i\le b}\,(\frac{(1+x_i)^2}{1-x_i^2}) < TOL \quad . \tag{1.4}$$

If we have no constraint on the condition numbers (i.e. $TOL=\infty$), then we have the following stronger test for stability:

$$\max_{1\le i\le b} \epsilon\, \frac{\sqrt{2}\cdot(p_i+q_i)}{\mathrm{Dif}_\lambda(\sigma_i,\sigma-\sigma_i)} < 1 \quad . \tag{1.5}$$

We will explain the quantities p_i, q_i, $\mathrm{Dif}_u(\sigma_i,\sigma-\sigma_i)$, $\mathrm{Dif}_l(\sigma_i,\sigma-\sigma_i)$, and $\mathrm{Dif}_\lambda(\sigma_i,\sigma-\sigma_i)$ in the next section. For now let us say that these bounds are fairly tight, and in fact the factor $b\cdot\max_{ij}\,(p_i,q_j)$ in the theorem is essentially the least possible value for the maximum of $\kappa(P)$ and $\kappa(Q)$ where P and Q block diagonalize $A-\lambda B$, the center of P(ϵ).

The quantities Dif_u, Dif_l, p_l, q_l, and Dif_λ may also be straightforwardly (if possibly expensively) computed using standard software packages. Also, nearly best conditioned block diagonalizing P and Q in (1.1) can also be computed. Therefore, it is possible to computationally verify conditions (1.3) to (1.5) and so to determine whether or not a decomposition is stable as defined above, as well as to compute the decomposition.

The criterion (1.3) is essentially due to Stewart [Stewart1]. Our contribution is the bound on $\kappa(P)$ and $\kappa(Q)$ in (1.4), as well as the stronger bound in (1.5).

The case when $A-\lambda B$ is singular is more difficult than the regular case because the eigenstructures we are interested in computing are nongeneric, i.e. they will be destroyed by almost all perturbations of the pencil $A-\lambda B$. The structure which interests us will exist only when A and B lie in a proper variety [Van Dooren1, Waterhouse]. A *proper variety* is the solution set of a set of polynomial equations in the entries of A and B such that the solution set is of positive codimension and hence of measure zero. What can stability mean in this context? For example, a nonsquare pencil $A-\lambda B$ will have the same KCF for almost all A and B, and this KCF can be determined by the dimensions of A and B alone. Also, it is possible to perturb a square singular pencil arbitrarily little making it regular with its eigenvalues anywhere in the extended complex plane [Wilkinson].

The known algorithms for computing nongeneric eigenspaces [Van Dooren1, Kågström1, Kågström2] are stable in the usual sense: they return the eigenstructure for a nongeneric pencil within a small distance of the input pencil, if one exists. Small perturbations in the input pencil (or during the computation) lead to the eigenstructure for a different nongeneric pencil, but, under certain assumptions, one with the same kind of eigenstructure (i.e. the same singular structure in the KCF) as the other nongeneric pencil. In other words, the algorithms perturb the input pencil in such a way that it is guaranteed to lie on the same proper variety. This leads us to ask whether we can deduce perturbation bounds of the features of the eigenstructure which remain stable under this kind of nongeneric perturbation. In order to do so, we must determine what features are stable and of interest. We choose features of interest in systems theory [Van Dooren2, Rosenbrock], which we describe below.

Care must be exercised in choosing stable features, since the following example shows that the spaces spanned by the P_l and Q_l are no longer all well defined in the singular case. Consider

$$P(A-\lambda B)Q^{-1} = \begin{bmatrix}1 & x\\ 0 & 1\end{bmatrix} \cdot \begin{bmatrix}\lambda & 1 & 0\\ 0 & 0 & \lambda-1\end{bmatrix} \cdot \begin{bmatrix}1 & 0 & x\\ 0 & 1 & x\\ 0 & 0 & 1\end{bmatrix}^{-1} = \begin{bmatrix}\lambda & 1 & 0\\ 0 & 0 & \lambda-1\end{bmatrix} .$$

As x grows large, the space spanned by Q_2 (the last column of Q) can become arbitrarily close to the space spanned by Q_1 (the first two columns of Q). Similarly the space spanned by P_2 (the last column of P) can become arbitrarily close to the space spanned by P_1 (the first column of P). Thus, we must modify the notion of deflating subspace used in the regular case, since these subspaces no longer all have unique definitions.

The correct concept to use is *reducing subspace*, as introduced in [Van Dooren3]. **P** and **Q** are reducing subspaces for $A-\lambda B$ if $\mathbf{P}=A\mathbf{Q}+B\mathbf{Q}$ and $\dim(\mathbf{P})=\dim(\mathbf{Q})-\dim(N_r)$, where N_r is the right null space of $A-\lambda B$ over the field of rational functions in λ. It is easy to express $\dim(N_r)$ in terms of the KCF of $A-\lambda B$: it is the number of L_k blocks in the KCF [Van Dooren3]. In the example above, N_r is one dimensional and spanned by $[1,-\lambda,0]^T$. The nontrivial pair of reducing subspaces are spanned by P_1 and Q_1 and are well defined.

We will prove that if a singular pencil is perturbed in such a way that it has reducing subspaces of the same dimension as the unperturbed pencil, then the perturbed subspaces will be close to the unperturbed spaces if the perturbation is small enough. Since a pencil can

have several reducing subspaces of the same dimension, care must be taken to distinguish these different subspaces; we show that reducing subspaces of the perturbed pencil must either be close to the unperturbed spaces or a bounded distance away. This result is stated in Theorem 6.

We can apply this result to compute perturbation bounds for standard algorithms for computing reducing subspaces. For small enough perturbations of the input data of one of these algorithms, we will show that the algorithm computes reducing subspaces a small angle away from the unperturbed spaces. This application is incorporated in Algorithm 1.

Reducing subspaces are of interest in systems theory, as the following example shows. Consider the pencil $A-\lambda B = [D|C-\lambda I]$. The pencil $[D|C-\lambda I]$ has the same KCF for nearly all matrices C and D of a fixed dimension: L_k blocks only. In systems theory [Wonham] this property is called *complete controllability* of the pair (C,D). Nonetheless, systems theory applications [Wonham, Van Dooren2] require knowing if (C,D) is uncontrollable ($[D|C-\lambda I]$ has a regular part in its KCF) or nearly so. If (C,D) is uncontrollable, we may ask what its smallest pair of reducing subspaces **P** and **Q** are. In systems theory the **P** of this pair is called the *controllable subspace* $\mathbf{C}(C,D)$ of the pair (C,D), and is of interest in designing control systems. Theorem 6 on the stability of reducing subspaces will imply that the controllable subspace of a pair is computed stably by the standard algorithms. This result is stated in Corollary 4: if the system (C,D) is perturbed to $(C+E_C,D+E_D)$ such that the controllable subspace of perturbed pencil $\mathbf{C}(C+E_C,D+E_D)$ has the same dimension as $\mathbf{C}(C,D)$, and if $\|(E_C,E_D)\|_E$ is small enough, then the largest angle between $\mathbf{C}(C,D)$ and $\mathbf{C}(C+E_C,D+E_D)$ is bounded by a constant (which can be computed using standard software from C and D) times $\|(E_C,E_D)\|_E$ (or else the largest angle is bounded away from 0). Similar comments apply to the unobservable subspace and other features of a control system (see [Van Dooren2] for a discussion).

Also of interest are the eigenvalues of the regular part of a nongeneric singular pencil. A generic perturbation of a square singular pencil will make the pencil regular and can be chosen to put the eigenvalues in arbitrary locations in the extended complex plane. A generic perturbation of a nonsquare singular pencil will make the regular part disappear. As before, however, the standard algorithms will produce a nongeneric pencil with a regular part of the same size as nearby pencils. We will show that the spectrum of the regular part is stable, i.e. we will derive perturbation bounds on the eigenvalues of the regular part.

Returning to our systems theory example, the eigenvalues of the regular part of the pencil $[D|C-\lambda I]$ are called the *input decoupling zeroes* or *uncontrollable modes* of the pair (C,D), and are of interest in designing control systems. They are the eigenvalues of a system $\dot{x} = Cx+Du$ which cannot be controlled by any choice of feedback $u = Fx$. Our result on the stability of the spectrum of the regular part will show that these eigenvalues are computed stably by the standard algorithms. This result is stated in Corollary 5.

All the results on the singular case as well as applications to systems theory are new.

The rest of this paper is organized as follows. Section 2 notation and basic lemmas. Section 3 will be devoted to analyzing the regular pencil case and proving Theorem A. This work is a generalization of work done in the doctoral thesis of one of the authors [Demmel2, Demmel and Kågström]. Section 4 will be devoted to the stability results for the singular case with applications of these results to problems in systems theory. Section 5 contains numerical examples.

2. Notation and Basic Lemmas.

$\|x\|$ will denote the Euclidean norm of the vector x. $\|A\|$ will denote the matrix norm induced by the Euclidean vector norm. $\|A\|_E$ will denote the Frobenius norm. $\sigma_{\min}(A)$ and $\sigma_{\max}(A)$ will denote the smallest and largest singular values of the matrix A. If $A=U\cdot\mathrm{diag}(\sigma_i)\cdot V^*$ is the singular value decomposition, then the pseudoinverse of A, A^+, is given by $A^+=V\cdot\mathrm{diag}(\sigma_i^+)\cdot U^*$, where $\sigma_i^+=\sigma_i^{-1}$ if $\sigma_i\neq 0$ and $\sigma_i^+=0$ if $\sigma_i=0$. $\kappa(A)$ will denote the condition number $\sigma_{\max}(A)/\sigma_{\min}(A)$ of the matrix A; this applies to nonsquare A as well. $A\otimes B$ will denote the Kronecker product of the two matrices A and B : $A\otimes B \equiv [A_{ij}B]$. Let col$A$ denote the column vector formed by taking the columns of A and stacking them atop one another from left to right. Thus if A is m by n, col$A$ is mn by 1 with its first m entries being column 1 of A, its second m entries being column 2 of A, and so on.

The following lemma is a generalization of a theorem in [Stewart1, Theorem 3.1] which we need for both the regular and singular cases later. Let T be an m by n matrix of full rank, and T^+ its n by m pseudoinverse. Let ϕ be a continuous map from $\mathbf{C}^m$ to $\mathbf{C}^n$ satisfying

i) $\|\phi(x)\| \le \|\phi\| \cdot \|x\|^2$

ii) $\|\phi(x)-\phi(y)\| \le 2\cdot\|\phi\| \max(\|x\|, \|y\|)\cdot\|x-y\|$

for some constant $\|\phi\|>0$. Let g be in $\mathbf{C}^n$.

Consider the equations

$$Tx = g-\phi(x) \tag{2.1}$$

and

$$x = T^+(g-\phi(x)) \tag{2.2}$$

We are interested in whether these equations have a solution, and what the solutions have to do with one another. There are three cases, depending on whether $m=n$ (T is invertible), $m<n$ ((2.1) is underdetermined), or $m>n$ ((2.1) is overdetermined). The case $m=n$ is dealt with by Stewart's original theorem and will be used to analyze the regular pencil case. The other two cases are not dealt with by Stewart and are used in the singular case.

Lemma 1: Assume that $\|T^+\|$, $\|\phi\|$, and $\|g\|$ satisfy

$$\kappa \equiv \|g\|\cdot\|\phi\|\cdot\|T^+\|^2 < \frac{1}{4} .$$

Then equation (2.2) has a unique solution $\tilde{x}$ inside the ball

$$\|\tilde{x}\| \le \frac{1-(1-4\kappa)^{1/2}}{2\kappa}\cdot\|g\|\cdot\|T^+\| < 2\cdot\|g\|\cdot\|T^+\| . \tag{2.3}$$

This solution $\tilde{x}$ of (2.2) has the following relationship to the solution $\hat{x}$ of (2.1):

Case 1: If $m=n$, $\hat{x}=\tilde{x}$ is the unique solution of (2.1) (in the ball).

Case 2: If $m<n$, $\hat{x}=\tilde{x}$ is a solution of (2.1), but it is not necessarily unique.

Case 3: If $m>n$, and if (2.1) has a solution $\hat{x}$ in the ball, then $\hat{x}=\tilde{x}$. Thus, (2.2) may have a solution whereas (2.1) may not.

Furthermore, in cases 1 and 3, if (2.1) has a solution which does not lie inside the ball in (2.3), it must lie outside the ball:

$$\|\hat{x}\| \ge \frac{1+(1-4\kappa)^{1/2}}{2\cdot\|T^+\|\cdot\|\phi\|} . \tag{2.4}$$

Proof: The proof that (2.2) has a solution under the given conditions is identical to the original proof of Stewart's theorem, so we will just outline it here. Let $\kappa = \|g\| \cdot \|\phi\| \cdot \|T^+\|^2 < 1/4$. Define the iteration

$$x_{i+1} = T^+(g - \phi(x_i))$$

with $x_0 = 0$. It is easy to show this is a contraction and converges to a unique solution $\hat{x}$ satisfying

$$\|\hat{x}\| \le \frac{1-(1-4\kappa)^{1/2}}{2\kappa} \cdot \|g\| \cdot \|T^+\| < 2\|g\| \cdot \|T^+\| \ .$$

Given this solution for (2.2) the other results are easy. Case 1 is the standard case already considered in [Stewart1]. In this case, $T^+ = T^{-1}$ so equations (2.1) and (2.2) are equivalent. In case 2, note that $T \cdot T^+ = I_m$, so multiplying $\hat{x} = T^+(g - \phi(\hat{x}))$ on the left by T yields the result. Nonuniqueness is easy to see: consider $\phi \equiv 0$. In case 3 we have $T^+ \cdot T = I_n$, so if $\hat{x}$ is a solution of (2.1), multiplying (2.1) on the left by T^+ yields $\hat{x} = T^+(g - \phi(\hat{x}))$, implying $\hat{x} = \hat{x}$.

The proof of (2.4) is as follows. Given a solution $\hat{x}$ of (2.1) it must satisfy

$$\frac{\|\hat{x}\|}{\|T^+\|} \ge \|T\hat{x}\| = \|g + \phi(\hat{x})\| \le \|g\| + \|\phi\| \cdot \|\hat{x}\|^2 \ .$$

Solving the quadratic inequality

$$\frac{\|\hat{x}\|}{\|T^+\|} \le \|g\| + \|\phi\| \cdot \|\hat{x}\|^2 \ .$$

for $\|\hat{x}\|$ yields two inequalities for $\|\hat{x}\|$, the one in (2.3) and the one in (2.4). Q.E.D.

Note that this theorem works in real vector spaces as well.

We need another lemma from the literature which we cite here. Let $P = [P_1 | \cdots | P_b]$ be a square partitioned matrix where P_i has n_i columns. We want to know how well conditioned we can make P subject to the constraint that the columns of P_i span a given n_i-dimensional subspace $\mathbf{P}_i$. Clearly, if P satisfies this constraint then any other matrix also satisfying the constraint can be written as PD, where $D = \mathrm{diag}(D_{11}, \ldots, D_{nn})$, D_{ii} a nonsingular n_i by n_i matrix.

Lemma 2: [Demmel1] Let θ_i be the smallest angle between any nonzero vector in $\mathbf{P}_i$ and the the subspace spanned by all the other $\mathbf{P}_j$, $j \ne i$. Then if P is any matrix such that P_i spans $\mathbf{P}_i$, we have

$$\max_{1 \le i \le b} \cot \theta_i/2 \le \inf_D \kappa(PD) \le b \cdot \max_{1 \le i \le b} \csc \theta_i \ . \tag{2.5}$$

Another way to express this inequality is as follows: partition P^{-1} into groups of rows as follows:

$$P^{-1} = \begin{bmatrix} P^1 \\ \cdot \\ P_b \end{bmatrix}$$

where P^i is n_i by n. Then

$$\csc \theta_i = \|P_i \cdot P^i\|$$

and

$$\cot\ \theta_i/2 = \|P_i \cdot P^i\| + (\|P_i \cdot P^i\|^2 - 1)^{1/2} \ .$$

Thus, the upper and lower bounds in (2.5) differ by a factor of at most b. The matrices $P_i \cdot P^i$ are oblique projections onto $\mathbf{P}_i$ parallel to $\mathbf{P}_j$, $i \neq j$. If we choose D so that the columns of $P_i D_{ii}$ are orthonormal, then $\kappa(PD)$ lies within the bounds of (2.5).

If $b=2$ and without loss of generality we assume

$$P = \begin{bmatrix} I_{n_1} & R \\ 0 & I_{n_2} \end{bmatrix}$$

then the choice $D_{11} = I_{n_1}$ and $D_{22} = (1 + \|R\|^2)^{-1/2} I_{n_2}$ makes $\kappa(KD)$ achieve its lower bound above, namely

$$\kappa(PD) = \|R\| + (1 + \|R\|^2)^{1/2}$$

In terms of R, we may write

$$\|P_i\, P^i\| = (1 + \|R\|^2)^{1/2} \ .$$

3. Regular Pencils

In trying to stably decompose $A-\lambda B$ into b blocks as in (1.1), we will first study decomposing $A-\lambda B$ into two blocks. Let $\sigma=\sigma_1 \bigcup \sigma_2$ be a possible decomposition of $\sigma(A-\lambda B)$ into disjoint subsets. Define the *dissociation* of σ_1 and σ_2, written diss(σ_1 , σ_2), as the smallest perturbation $(A+E)-\lambda(B+F)$ of $A-\lambda B$ (measured as $\|(E,F)\|_E$) that makes an eigenvalue $\lambda_1 \in \sigma_1$ coalesce with $\lambda_2 \in \sigma_2$. (Think of λ_1 and λ_2 as continuous function of $0 \le x \le 1$, where $\lambda_i(x)$ is an eigenvalue of $(A+xE)-\lambda(B+xF)$ and $\lambda_i(0)=\lambda_i$. As x increases from 0 to 1, $\lambda_1(x)$ and $\lambda_2(x)$ move continuously until possibly $\lambda_1(1)=\lambda_2(1)$. This is possible for $\|(E,F)\| \ge$ diss(σ_1 , σ_2) but no smaller.) It is easy to see that the first condition of a stable decomposition of P(ϵ), that the number of eigenvalues in each σ_i remain constant, holds if and only if diss(σ_i , $\sigma-\sigma_i$)$>\epsilon$ for all i. Thus, if we can compute lower bounds on diss(σ_i , $\sigma-\sigma_i$), a decomposition of P(ϵ) will satisfy the first criterion of stability if these lower bounds are all greater than ϵ. This will be our approach for proving Theorem A.

We begin with a simple lower bound on diss(σ_i , $\sigma-\sigma_i$) based on an extension of the Bauer-Fike Theorem [Bauer and Fike] to pencils. It can also be seen as an extension of Gershgorin's Theorem to pencils [Stewart2]. We include it even though it is generally weaker than our later results because it illustrates the nature of the problem.

Lemma 3: Suppose $P^{-1}(A-\lambda B)Q = \mathrm{diag}(\alpha_i-\lambda\beta_i)$. Then if s/c (where $|s|^2+|c|^2=1$) is an eigenvalue of the perturbed pencil $(A+E)-\lambda(B+F)$, we have

$$\|P^{-1}\| \cdot \|Q\| \cdot \|(E,F)\|_E \ge \min_i |c\alpha_i - s\beta_i|$$

Proof: Assume without loss of generality that s/c is not an eigenvalue of the unperturbed pencil. Then

$$\begin{aligned}
0 &= \det(c(A+E)-s(B+F)) \\
&= \det(P^{-1}(c(A+E)-s(B+F))Q) \\
&= \det(\mathrm{diag}(c\alpha_i-s\beta_i)\cdot(I+\mathrm{diag}((c\alpha_i-s\beta_i)^{-1})P^{-1}(cE-sF)Q)) \\
&= \det(I+\mathrm{diag}((c\alpha_i-s\beta_i)^{-1})P^{-1}(cE-sF)Q)
\end{aligned}$$

implying

$$\begin{aligned}
1 &\le \|\mathrm{diag}((c\alpha_i-s\beta_i)^{-1})P^{-1}(cE-sF)Q\|_E \\
&\le (\min_i |c\alpha_i-s\beta_i|)^{-1} \|P^{-1}\| \cdot \|Q\| \cdot \|cE-sF\|_E \\
&\le (\min_i |c\alpha_i-s\beta_i|)^{-1} \|P^{-1}\| \cdot \|Q\| \cdot \|(E,F)\|_E
\end{aligned}$$

which is equivalent to the result claimed. Q.E.D.

Stewart in [Stewart2] expresses this result in terms of an upper bound on the chordal metric

$$\chi(\lambda,\lambda') = \frac{|\lambda - \lambda'|}{(1+\lambda^2)^{1/2} \cdot (1+\lambda'^2)^{1/2}} ,$$

where λ is an original eigenvalue and λ' is a perturbed eigenvalue. We choose not to do so here because we do not need all the properties of a metric for our results.

Just as the Bauer-Fike theorem for the standard eigenproblem showed that the eigenvalues of the perturbed matrix must lie in disks centered at the eigenvalues of the unperturbed matrix, this lemma shows that the eigenvalues of the perturbed pencil lie in

certain regions around the eigenvalues of the unperturbed pencil. By requiring these regions to be disjoint, we get a lower bound on diss(σ_1, σ_2).

Theorem 1: Suppose $P^{-1}(A-\lambda B)Q = \mathrm{diag}(\alpha_{1l}-\lambda\beta_{1l}, \alpha_{2j}-\lambda\beta_{2j})$ where $\alpha_{1l}/\beta_{1l} \in \sigma_1$ and $\alpha_{2j}/\beta_{2j} \in \sigma_2$. Then

$$\mathrm{diss}(\sigma_1, \sigma_2) \geq \frac{1}{\sqrt{2}\cdot\|P^{-1}\|\cdot\|Q\|}\cdot\min_{ij}\sigma_{\min}\begin{bmatrix}\alpha_{1i} & \alpha_{2j}\\ \beta_{1i} & \beta_{2j}\end{bmatrix}$$

Proof: From Lemma 3 we see that x can only exceed diss(σ_1, σ_2) if there exist i, j, c and s such that $|c|^2+|s|^2=1$ all satisfying

$$\|P^{-1}\|\cdot\|Q\|\cdot x \geq \max\left(|c\alpha_{1i}-s\beta_{1i}|, |c\alpha_{2j}-s\beta_{2j}|\right)$$

$$\geq \frac{1}{\sqrt{2}}\sigma_{\min}\left(\begin{bmatrix}\alpha_{1i} & \alpha_{2j}\\ \beta_{1i} & \beta_{2j}\end{bmatrix}\right).$$

In other words, $x \geq \mathrm{diss}(\sigma_1, \sigma_2)$ only if

$$x \geq \frac{1}{\sqrt{2}\cdot\|P^{-1}\|\cdot\|Q\|}\min_{ij}\sigma_{\min}\begin{bmatrix}\alpha_{1i} & \alpha_{2j}\\ \beta_{1i} & \beta_{2j}\end{bmatrix},$$

which implies the result. Q.E.D.

Unfortunately, $\|P^{-1}\|\cdot\|Q\|$ may be quite large and the lower bound of this theorem quite tiny because $A-\lambda B$ may be hard to diagonalize, even if σ_1 and σ_2 are well separated. A more refined lower bound on diss(σ_1, σ_2) which takes this into account is derived as follows.

We begin by reducing the pencil $A-\lambda B$ to a canonical form by a transformation which does not change distances or angles between subspaces, quantities we need to preserve. We quote the following lemma:

Lemma 4: [Stewart3] Any regular $A-\lambda B$ can be transformed into upper triangular form by multiplication on the left and right by unitary matrices. Further, the unitary matrices may be chosen so that the eigenvalues appear on the diagonal of the transformed pencil in any desired order.

Standard software (the QZ algorithm [Moler and Stewart] and EXCHQZ [Van Dooren4]) is available to compute this decomposition.

Suppose now without loss of generality that our original pencil is in upper triangular form

$$A-\lambda B = \begin{bmatrix}A_{11} & A_{12}\\ 0 & A_{22}\end{bmatrix} - \lambda\begin{bmatrix}B_{11} & B_{12}\\ 0 & B_{22}\end{bmatrix}$$

with $\dim(A_{ii}-\lambda B_{ii}) = n_i$ and $\sigma(A_{ii}-\lambda B_{ii}) = \sigma_i$, with σ_1 and σ_2 disjoint.

We next need to compute the deflating subspaces of this pencil belonging to σ_1 and σ_2. Equivalently, we want to find P and Q such that

$$P^{-1}(A-\lambda B)Q = \begin{bmatrix}A_{11} & 0\\ 0 & A_{22}\end{bmatrix} - \lambda\begin{bmatrix}B_{11} & 0\\ 0 & B_{22}\end{bmatrix} \tag{3.1}$$

with $\sigma(A_{ii}-\lambda B_{ii}) = \sigma_i$. It is easy to see that the pair of deflating subspaces belonging to σ_1 are both spanned by $P_1 = Q_1 = [I_{n_1}|0]^T$. Without loss of generality we seek the other deflating pair in the form $P_2 = [L^T|I_{n_2}]^T$ and $Q_2 = [R^T|I_{n_2}]^T$, which leads to the equation

$$\begin{bmatrix} I_{n_1} & -L \\ 0 & I_{n_2} \end{bmatrix} \cdot \begin{bmatrix} A_{11}-\lambda B_{11} & A_{12}-\lambda B_{12} \\ 0 & A_{22}-\lambda B_{22} \end{bmatrix} \cdot \begin{bmatrix} I_{n_1} & R \\ 0 & I_{n_2} \end{bmatrix} = \begin{bmatrix} A_{11}-\lambda B_{11} & 0 \\ 0 & A_{22}-\lambda B_{22} \end{bmatrix}$$

or

$$A_{11}R-LA_{22} = -A_{12} \tag{3.2a}$$

$$B_{11}R-LB_{22} = -B_{12} \tag{3.2b}$$

This is a set of $2n_1n_2$ equations in $2n_1n_2$ unknowns, the entries of L and R. It is not difficult to see that we can rewrite (3.2a) and (3.2b) as follows:

$$\begin{bmatrix} I_{n_2}\otimes A_{11} & -A_{22}^T\otimes I_{n_1} \\ I_{n_2}\otimes B_{11} & -B_{22}^T\otimes I_{n_1} \end{bmatrix} \cdot \begin{bmatrix} \mathrm{col}R \\ \mathrm{col}L \end{bmatrix} \equiv Z_u \cdot \begin{bmatrix} \mathrm{col}R \\ \mathrm{col}L \end{bmatrix} = \begin{bmatrix} -\mathrm{col}A_{12} \\ -\mathrm{col}B_{12} \end{bmatrix} \tag{3.3}$$

One can show Z_u is nonsingular if and only if σ_1 and σ_2 are disjoint [Stewart3] as we have assumed. Clearly, If Z_u is ill conditioned, L and R may be very large, implying the deflating subspaces for σ_1 and σ_2 are close together. As we will see later, this means we can expect diss(σ_1 , σ_2) to be small as well. With this motivation, we define as in [Stewart1]

$$\mathrm{Dif}_u(A_{11},A_{22};B_{11},B_{22}) \equiv \sigma_{\min}(Z_u) \quad ,$$

the smallest singular value of Z_u. A trivial consequence of this definition is that

$$\| (L,R)\|_E \le \| (A_{12},B_{12})\|_E \,/\, \mathrm{Dif}_u(A_{11},A_{22};B_{11},B_{22}) \quad .$$

It is easy to see that Dif_u is not changed if we multiply A_{ii} and B_{ii} on the right by any unitary U_i and on the left by any unitary V_i. This implies that given $A-\lambda B$, Dif_u is determined only by specifying $\sigma_1=\sigma(A_{11}-\lambda B_{11})$ and $\sigma_2=(A_{22}-\lambda B_{22})$, in that order. (We will return to the dependence on order later.) We will therefore write $\mathrm{Dif}_u(\sigma_1,\sigma_2)$ when $A-\lambda B$ is clear from context, or just Dif_u if σ_1 and σ_2 are clear as well. We record a fact about Dif_u we will need later.

Lemma 5: [Stewart1]

$\mathrm{Dif}_u(A_{11}+E_{11},A_{22}+E_{22};B_{11}+F_{11},B_{22}+F_{22}) \ge \mathrm{Dif}_u(A_{11},A_{22};B_{11},F_{22}) - \sqrt{2}\cdot \| (E_{11},E_{22},F_{11},F_{22})\|_E$

Other properties of Dif_u can also be found in [Stewart1].

So now we know how to compute L and R, and therefore blockdiagonalizing P and Q for $A-\lambda B$. For our next lower bound on diss(σ_1 , σ_2), we will need to use slightly different P and Q. Just as Dif_u is specified by σ_1 and σ_2 alone, it is easy to see that the norms of L and R are determined only by σ_1 and σ_2. Let p denote $(1+\|L\|^2)^{1/2}$ and q denote $(1+\|R\|^2)^{1/2}$. From Lemma 2, we know p and q are lower bounds on the condition numbers of any blockdiagonalizing P and Q, respectively. In the language of Lemma 2, p is the norm of the projector onto either left deflating subspace parallel to the other. Similarly, q is the norm of a projector on the right. p and q are defined given only σ_1 and σ_2 since $\|L\|$ and $\|R\|$ are. Define also

$$P_0 = \begin{bmatrix} I_{n_1} & L \\ 0 & I_{n_2} \end{bmatrix} \cdot \begin{bmatrix} p^{1/2} I_{n_1} & 0 \\ 0 & q^{-1/2} I_{n_2} \end{bmatrix}$$

and

$$Q_0 = \begin{bmatrix} I_{n_1} & R \\ 0 & I_{n_2} \end{bmatrix} \cdot \begin{bmatrix} p^{1/2} I_{n_1} & 0 \\ 0 & q^{-1/2} I_{n_2} \end{bmatrix} .$$

It is easy to verify that

$$\|P_0\| \cdot \|Q_0^{-1}\| = (\kappa(P_0) \cdot \kappa(Q_0))^{1/2} \le p + q \quad . \tag{3.4}$$

The quantity $\|P_0\| \cdot \|Q_0^{-1}\|$ will play the same role for regular pencils as the condition number of the best conditioned block diagonalizing similarity plays for the standard eigenvalue problem: they measure the sensitivity of eigenvalues to perturbations.

We need one more definition before presenting our next lower bound on diss(σ_1 , σ_2):

$$\mathrm{Dif}_\lambda(A_{11},A_{22};B_{11},B_{22}) \equiv \inf_{\substack{c,s \\ |c|^2+|s|^2=1}} (\sigma^2_{\min}(cA_{11}-sB_{11}) + \sigma^2_{\min}(cA_{22}-sB_{22}))^{1/2} \; .$$

Just as with Dif_u it turns out Dif_λ is specified by $A-\lambda B$ and σ_1 and σ_2 alone, permitting us to write $\mathrm{Dif}_\lambda(\sigma_1,\sigma_2)$ when $A-\lambda B$ is clear from context and just Dif_λ when σ_1 and σ_2 are clear as well.

Dif_λ will play the same role in this analysis as sep_λ [Varah, Demmel2] played for the standard eigenvalue problem: it is the size of the smallest perturbation that makes the two pencils $A_{11}-\lambda B_{11}$ and $A_{22}-\lambda B_{22}$ have a common eigenvalue:

Lemma 6: Let $\|(E_{11},E_{22},F_{11},F_{22})\|_E$ be the size of the smallest perturbation such that $(A_{11}+E_{11})-\lambda(B_{11}+F_{11})$ and $(A_{22}+E_{22})-\lambda(B_{22}+F_{22})$ have a common eigenvalue. Then

$$\mathrm{Dif}_\lambda(A_{11},A_{22};B_{11},B_{22}) = \|(E_{11},E_{22},F_{11},F_{22})\|_E \; .$$

Proof: Choose c and s in the definition of Dif_λ to attain the infimum. Next choose E and F each of rank 1 and such that $\|E\|_E^2+\|F\|_E^2 = \mathrm{Dif}_\lambda^2$ satisfying $\sigma_{\min}(cA_{11}-sB_{11}+E)=0$ and $\sigma_{\min}(cA_{22}-sB_{22}+F)=0$. These last equations can be rewritten

$$\sigma_{\min}(c(A_{11}+\bar{c}E)-s(B_{11}-\bar{s}E)) \equiv \sigma_{\min}(c(A_{11}+E_{11})-s(B_{11}+F_{11})) = 0$$

and

$$\sigma_{\min}(c(A_{22}+\bar{c}F)-s(B_{22}-\bar{s}F)) \equiv \sigma_{\min}(c(A_{22}+E_{22})-s(B_{22}+F_{22})) = 0$$

Now

$$\|(E_{11},E_{22},F_{11},F_{22})\|_E^2 = \|E\|_E^2 + \|F\|_E^2 = \mathrm{Dif}_\lambda^2 \;\; ,$$

i.e. this particular choice of E_{11}, E_{22}, F_{11}, and F_{22} satisfies the upper bound, and hence so must the smallest perturbation. To show Dif_λ is a lower bound, let E_{ii} and F_{ii} for $i=1,2$ be any perturbations such that $(A_{11}+E_{11})-\lambda(B_{11}+F_{11})$ and $(A_{22}+E_{22})-\lambda(B_{22}+F_{22})$ have a common eigenvalue s/c with $|c|^2+|s|^2=1$. Then

$$0=\sigma_{\min}(c(A_{ii}+E_{ii})-s(B_{ii}+F_{ii})) = \sigma_{\min}(cA_{ii}-sB_{ii}+(cE_{ii}-sF_{ii})) \ge \sigma_{\min}(cA_{ii}-sB_{ii}) - \|cE_{ii}-sF_{ii}\|_E$$

so by the definition of Dif_λ we get

$$\mathrm{Dif}_\lambda \le (\|cE_{11}-sF_{11}\|_E^2 + \|cE_{22}-sF_{22}\|_E^2)^{1/2} \le \|(E_{11},E_{22},F_{11},F_{22})\|_E \; .$$

Q.E.D.

This lemma immediately yields a simple upper bound on diss(σ_1 , σ_2):

Corollary 1:

$$\mathrm{diss}(\sigma_1 , \sigma_2) \le \mathrm{Dif}_\lambda(\sigma_1,\sigma_2) \ .$$

We are now ready to prove our next lower bound on $\mathrm{diss}(\sigma_1 , \sigma_2)$:

Theorem 2:

$$\mathrm{diss}(\sigma_1 , \sigma_2) \ge \frac{\mathrm{Dif}_\lambda(\sigma_1,\sigma_2)}{\sqrt{2}\cdot(p+q)} \ .$$

Proof: Choose c and s so that $|c|^2+|s|^2=1$ and also so that s/c is an eigenvalue of the perturbed pencil $(A+E)-\lambda(B+F)$ but not of $A-\lambda B$. Then

$$\begin{aligned} 0 &= \det(c(A+E)-s(B+F)) \\ &= \det(P_0^{-1}(cA-sB)Q_0+P_0^{-1}(cE-sF)Q_0) \\ &= \det(\begin{bmatrix} cA_{11}-sB_{11} & 0 \\ 0 & cA_{22}-sB_{22} \end{bmatrix} + P_0^{-1}(cE-sF)Q_0) \\ &= \det(I + \begin{bmatrix} cA_{11}-sB_{11} & 0 \\ 0 & cA_{22}-sB_{22} \end{bmatrix}^{-1} P_0^{-1}(cE-sF)Q_0) \end{aligned}$$

implying

$$\begin{aligned} 1 &\le \left\| \begin{bmatrix} cA_{11}-sB_{11} & 0 \\ 0 & cA_{22}-sB_{22} \end{bmatrix}^{-1} P_0^{-1}(cE-sF)Q_0 \right\|_E \\ &\le \max\left(\|(cA_{11}-sB_{11})^{-1}\|, \|(cA_{22}-sB_{22})^{-1}\|\right) \cdot \|P_0^{-1}\| \cdot \|Q_0\| \cdot \|cE-sF\|_E \end{aligned}$$

or, rearranging and using (3.4)

$$\frac{\min\left(\|(cA_{11}-sB_{11})^{-1}\|^{-1}, \|(cA_{22}-sB_{22})^{-1}\|^{-1}\right)}{p+q} \le \|cE-sF\|_E \ .$$

This in turn implies

$$\|(E,F)\|_E \ge \|cE-sF\|_E \ge \frac{\min\left(\sigma_{\min}(cA_{11}-sB_{11}), \sigma_{\min}(cA_{22}-sB_{22})\right)}{p+q} \ .$$

Thus, the eigenvalues s/c of the perturbed pencil lie in clusters about the eigenvalues of the unperturbed one, these clusters being defined by

$$\|(E,F)\|_E \ge \frac{\sigma_{\min}(cA_{ii}-sB_{ii})}{p+q}$$

for $i=1,2$. These clusters can only overlap (a necessary condition for coalescence of eigenvalues) if for some s and c

$$\begin{aligned} \|(E,F)\|_E &\ge \frac{\max\left(\sigma_{\min}(cA_{11}-sB_{11}), \sigma_{\min}(cA_{22}-sB_{22})\right)}{p+q} \\ &\ge \frac{\mathrm{Dif}_\lambda(\sigma_1,\sigma_2)}{\sqrt{2}\cdot(p+q)} \end{aligned}$$

which implies our result. Q.E.D.

An immediate corollary of our last theorem is

Corollary 2: Suppose

$$A-\lambda B = \begin{bmatrix} A_{11}-\lambda B_{11} & 0 \\ 0 & A_{22}-\lambda B_{22} \end{bmatrix}$$

is block diagonal. Then

$$\mathrm{Dif}_\lambda(\sigma_1,\sigma_2) \geq \mathrm{diss}(\sigma_1\ ,\ \sigma_2) \geq \frac{1}{\sqrt{2}} \cdot \mathrm{Dif}_\lambda(\sigma_1,\sigma_2) \ .$$

Finally, we use Theorem 2 to prove part of Theorem A: deciding if a decomposition is stable if there is no constraint on the condition number of P and Q:

Corollary 3: Let $\sigma=\bigcup_{i=1}^{b} \sigma_i$ be a partitioning of the spectrum of $A-\lambda B$. Let p_i and q_i be the values of $(1+\|R\|^2)^{1/2}$ and $(1+\|L\|^2)^{1/2}$ corresponding to σ_i and $\sigma-\sigma_i$ in (3.3). Then $\bigcup_{i=1}^{b} \sigma_i$ is a stable decomposition of $\mathrm{P}(\epsilon)$ in (1.1) if there is no constraint TOL on the condition numbers of P and Q and if

$$\max_{1\leq i\leq b} \epsilon \cdot \frac{\sqrt{2}\cdot(p_i+q_i)}{\mathrm{Dif}_\lambda(\sigma_i,\sigma-\sigma_i)} < 1 \ .$$

Proof: The last inequality implies that $\epsilon < \mathrm{diss}(\sigma_i\ ,\ \sigma-\sigma_i)$ for all $1\leq i\leq b$. Q.E.D.

We can now derive our third lower bound on $\mathrm{diss}(\sigma_1\ ,\ \sigma_2)$, one which will also lead to bounds on $\kappa(P)$ and $\kappa(Q)$, where P and Q are the blockdiagonalizing equivalence transformations in (1.1). We will use essentially the same approach as in [Stewart1]. We assume as before that $A-\lambda B$ is upper triangular and that P and Q blockdiagonalize it as in equation (3.1). Consider the perturbed pencil $(A+E)-\lambda(B+F)$. Premultiplying by P^{-1} and postmultiplying by Q yields the pencil

$$P^{-1}\cdot((A+E)-\lambda(B+F))\cdot Q = \begin{bmatrix} A_{11}+E_{11} & E_{12} \\ E_{21} & A_{22}+E_{22} \end{bmatrix} - \lambda \begin{bmatrix} B_{11}+F_{11} & F_{12} \\ F_{21} & B_{22}+F_{22} \end{bmatrix} \tag{3.5}$$

We now seek P_{EF}^{-1} and Q_{EF} of the forms

$$P_{EF}^{-1} = \begin{bmatrix} I_{n_1} & -L_1 \\ L_2 & I_{n_2} \end{bmatrix}$$

and

$$Q_{EF} = \begin{bmatrix} I_{n_1} & R_1 \\ -R_2 & I_{n_2} \end{bmatrix} \ .$$

such that premultiplying (3.5) by P_{EF}^{-1} and postmultiplying it by Q_{EF} blockdiagonalizes it. Performing these multiplications and rearranging the result yields

$$\begin{bmatrix} (A_{11}+E_{11}-L_1E_{21})\,(I+R_1R_2) & (A_{11}+E_{11})R_1-L_1(A_{22}+E_{22}) + E_{12} - L_1E_{22}R_1 \\ L_2(A_{11}+E_{11})-(A_{22}+E_{22})R_2 + E_{21} - L_2E_{12}R_2 & (A_{22}+E_{22}+L_2E_{12})\,(I+R_2R_1) \end{bmatrix} \tag{3.6}$$

$$-\lambda\begin{bmatrix} (B_{11}+F_{11}-L_1F_{21})\,(I+R_1R_2) & (B_{11}+F_{11})R_1-L_1(B_{22}+F_{22}) + F_{12} - L_1F_{21}R_1 \\ L_2(B_{11}+F_{11})-(B_{22}+F_{22})R_2 + F_{21} - L_2F_{12}R_2 & (B_{22}+F_{22}+L_2F_{12})\,(I+R_2R_1) \end{bmatrix} .$$

Setting the upper right and lower left corners of the pencil two zeroes yields two sets of equations:

$$(A_{11}+E_{11})R_1-L_1(A_{22}+E_{22}) = -E_{12}+L_1E_{21}R_1 \tag{3.7a}$$

$$(B_{11}+F_{11})R_1-L_1(B_{22}+F_{22}) = -F_{12}+L_1F_{21}R_1 \tag{3.7b}$$

and

$$L_2(A_{11}+E_{11})-(A_{22}+E_{22})R_2 = -E_{21}+L_2E_{12}R_2 \tag{3.8a}$$

$$L_2(B_{11}+F_{11})-(B_{22}+F_{22})R_2 = -F_{21}+L_2F_{12}R_2 \ . \tag{3.8b}$$

We wish to apply Lemma 1 to solve these sets of nonlinear equations. To solve (3.7) we make the identifications $x \equiv [L_1|R_1]$,

$$Tx \equiv \begin{bmatrix} (A_{11}+E_{11})R_1-L_1(A_{22}+E_{22}) \\ (B_{11}+F_{11})R_1-L_1(B_{22}+F_{22}) \end{bmatrix} , \tag{3.9a}$$

$$g \equiv \begin{bmatrix} -E_{12} \\ -F_{12} \end{bmatrix} , \tag{3.9b}$$

and

$$\phi(x) \equiv \begin{bmatrix} L_1E_{21}R_1 \\ L_1F_{21}R_1 \end{bmatrix} . \tag{3.9c}$$

From Lemma 5, we get

$$\|T^{+}\|^{-1} = \|T^{-1}\|^{-1} \geq \mathrm{Dif}_u(\sigma_1,\sigma_2) - \sqrt{2}\cdot\|(E_{11},E_{22},F_{11},F_{22})\|_E \ . \tag{3.9d}$$

To solve (3.8) we make the identifications $x \equiv [L_2|R_2]$,

$$Tx \equiv \begin{bmatrix} L_2(A_{11}+E_{11})-(A_{22}+E_{22})R_2 \\ L_2(B_{11}+F_{11})-(B_{22}+F_{22})R_2 \end{bmatrix} , \tag{3.10a}$$

$$g \equiv \begin{bmatrix} -E_{21} \\ -F_{21} \end{bmatrix} , \tag{3.10b}$$

and

$$\phi(x) \equiv \begin{bmatrix} L_2E_{12}R_2 \\ L_2F_{12}R_2 \end{bmatrix} . \tag{3.10c}$$

Using Kronecker products to express the linear operator T in (3.10a) yields

$$\begin{bmatrix} (A_{11}+E_{11})^T\otimes I_{n_2} & -I_{n_1}\otimes(A_{22}+E_{22}) \\ (B_{11}+F_{11})^T\otimes I_{n_2} & -I_{n_1}\otimes(B_{22}+F_{22}) \end{bmatrix} . \tag{3.11}$$

Swapping the first n_1n_2 columns with the last n_1n_2 columns and negating the whole matrix, none of which changes its singular values, yields

$$\begin{bmatrix} I_{n_1} \otimes (A_{22}+E_{22}) & -(A_{11}+E_{11})^T \otimes I_{n_2} \\ I_{n_1} \otimes (B_{22}+F_{22}) & -(B_{11}+F_{11})^T \otimes I_{n_2} \end{bmatrix}$$

which we recognize as the matrix whose smallest singular value is defined as

$$\mathrm{Dif}_u(A_{22}+E_{22},A_{11}+E_{11};B_{22}+F_{22},B_{11}+F_{11}) \geq \mathrm{Dif}_u(A_{22},A_{11};B_{22},B_{11}) - \sqrt{2} \cdot \| (E_{11},E_{22},F_{11},F_{22}) \|_E \ .$$

The quantity $\mathrm{Dif}_u(A_{22},A_{11};B_{22},B_{11})$ does not generally equal $\mathrm{Dif}_u(A_{11},A_{22};B_{11},B_{22})$ (unless the A_{ii} and B_{ii} are symmetric). In the interests of retaining our coordinate free formulation of our bounds, we therefore define

$$\mathrm{Dif}_l(\sigma_1,\sigma_2) \equiv \mathrm{Dif}_u(A_{22},A_{11};B_{22},B_{11})$$

where the fact that Dif_l depends only on σ_1 and σ_2 follows just as for Dif_u. This leads us to:

Theorem 3: Let p, q, Dif_u, and Dif_l be defined as above. Then

$$\mathrm{diss}(\sigma_1\ ,\ \sigma_2) \geq \frac{\min\ (\mathrm{Dif}_u(\sigma_1,\sigma_2)\ ,\ \mathrm{Dif}_l(\sigma_1,\sigma_2))}{(2\ (p^2 + q^2))^{1/2} + 2 \cdot \max\ (p,q)} \ .$$

Suppose that $\| (E,F) \|_E$ is less than this lower bound on $\mathrm{diss}(\sigma_1\ ,\ \sigma_2)$, and define

$$x \equiv \| (E,F) \|_E \cdot \frac{(2(p^2+q^2))^{1/2}+2 \cdot \max\ (p,q)}{\min\ (\mathrm{Dif}_u(\sigma_1,\sigma_2)\ ,\ \mathrm{Dif}_l(\sigma_1,\sigma_2))} < 1 \ .$$

Further define p_{EF} and q_{EF} to be the norms of the projectors onto the left and right deflating subspaces of the perturbed pencil $(A+E)-\lambda(B+F)$. Then

$$p_{EF} \leq 2 \cdot \frac{(1+x)^2}{1-x^2} \cdot p \quad \text{and} \quad q_{EF} \leq 2 \cdot \frac{(1+x)^2}{1-x^2} \cdot q \ .$$

Proof: We need to make two different choices of P and Q in (3.5). Both lead to the same lower bound on $\mathrm{diss}(\sigma_1\ ,\ \sigma_2)$, but only the first will lead to a bound on p_{EF}, and only the second to a bound on q_{EF}. The first choices of P and Q are

$$P = \begin{bmatrix} I_{n_1} & L \\ 0 & I_{n_2} \end{bmatrix} \cdot \begin{bmatrix} p^{1/2}\, I_{n_1} & 0 \\ 0 & p^{-1/2}\, I_{n_2} \end{bmatrix}$$

and

$$Q = \begin{bmatrix} I_{n_1} & R \\ 0 & I_{n_2} \end{bmatrix} \cdot \begin{bmatrix} p^{1/2}\, I_{n_1} & 0 \\ 0 & p^{-1/2}\, I_{n_2} \end{bmatrix} \ .$$

With this choice of P and Q, it is easy to verify that the E_{ij} in (3.6) satisfy $\| E_{11} \|_E \leq p \cdot \| E \|_E$, $\| E_{12} \|_E \leq q \cdot \| E \|_E$, $\| E_{21} \|_E \leq p \cdot \| E \|_E$, and $\| E_{22} \|_E \leq q \cdot \| E \|_E$. The $\| F_{ij} \|_E$ satisfy analogous inequalities.

Consider equation (3.7) and the corresponding identifications in (3.9). Substituting in these bounds on $\| E_{ij} \|_E$ and $\| F_{ij} \|_E$ yields:

$$\| T^{-1} \|^{-1} \geq \mathrm{Dif}_u(\sigma_1,\sigma_2) - (2 \cdot (p^2 + q^2))^{1/2} \cdot \| (E,F) \|_E \ ,$$

$$\| g \| \leq q \cdot \| (E,F) \|_E \ ,$$

and

$$\| \phi \| \leq p \cdot \| (E,F) \|_E \ .$$

From Lemma 1, we see that as long as

$$\frac{\|g\| \cdot \|\phi\|}{\|T^{-1}\|^{-2}} \le \frac{p \cdot q \cdot \|(E,F)\|^2}{(\mathrm{Dif}_u(\sigma_1,\sigma_2) - (2\,(p^2+q^2))^{1/2}\,\|(E,F)\|_E)^2} < \frac{1}{4} \ ,$$

or, solving for $\|(E,F)\|_E$,

$$\|(E,F)\|_E < \frac{\mathrm{Dif}_u(\sigma_1,\sigma_2)}{(2\,(p^2+q^2))^{1/2} + 2\,(p \cdot q)^{1/2}} \ , \tag{3.12}$$

then we can solve equations (3.7) for L_1 and R_1. Furthermore, assuming $\|(E,F)\|_E$ is less that the lower bound on diss$(\sigma_1\ ,\ \sigma_2)$ in the statement of the theorem, we see from Lemma 1 that

$$\begin{aligned}
\|(L_1,R_1)\|_E &< 2 \cdot \|g\| \cdot \|T^{-1}\| \\
&\le \frac{2\,q\ \|(E,F)\|_E}{\mathrm{Dif}_u(\sigma_1,\sigma_2) - (2 \cdot (p^2+q^2))^{1/2} \cdot \|(E,F)\|_E} \\
&\le \frac{\dfrac{x \cdot 2 \cdot q \cdot \mathrm{Dif}_u(\sigma_1,\sigma_2)}{(2 \cdot (p^2+q^2))^{1/2} + 2\max(p,q)}}{\mathrm{Dif}_u(\sigma_1,\sigma_2) - (2 \cdot (p^2+q^2))^{1/2} \cdot \dfrac{\mathrm{Dif}_u(\sigma_1,\sigma_2)}{(2 \cdot (p^2+q^2))^{1/2} + 2\max(p,q)}} \\
&\le \frac{x \cdot q}{\max\,(p,q)} \le x < 1 \ .
\end{aligned}$$

Now consider equation (3.8) and the corresponding identifications in (3.10). Substituting in these bounds on $\|E_{ij}\|_E$ yields:

$$\|T^{-1}\|^{-1} \ge \mathrm{Dif}_l(\sigma_1,\sigma_2) - (2 \cdot (p^2+q^2))^{1/2} \cdot \|(E,F)\|_E \ ,$$

$$\|g\| \le p \cdot \|(E,F)\|_E \ ,$$

and

$$\|\phi\| \le q \cdot \|(E,F)\|_E \ .$$

Using Lemma 1 as before, we see that if

$$\|(E,F)\|_E < \frac{\mathrm{Dif}_l(\sigma_1,\sigma_2)}{(2\,(p^2+q^2))^{1/2} + 2\,(p \cdot q)^{1/2}} \ , \tag{3.13}$$

then we can solve equations (3.8) for L_2 and R_2. We can also bound $\|(L_2,R_2)\|_E$ as above yielding

$$\|(L_2,R_2)\|_E < \frac{x \cdot p}{\max\,(p,q)} \le x < 1 \ .$$

This implies that P_{EF} and Q_{EF} are invertible, since if

$$P_{EF}^{-1} = \begin{bmatrix} I_{n_1} & -L_1 \\ L_2 & I_{n_2} \end{bmatrix}$$

then P_{EF} must equal

$$P_{EF} = \begin{bmatrix} I_{n_1} & L_1 \\ -L_2 & I_{n_2} \end{bmatrix} \cdot \begin{bmatrix} (I_{n_1} + L_1 L_2)^{-1} & 0 \\ 0 & (I_{n_2} + L_2 L_1)^{-1} \end{bmatrix}$$

which clearly exists since $\|L_i\| < 1$ for $i=1,2$. Similarly,

$$Q_{EF}^{-1} = \begin{bmatrix} I_{n_1} & -R_1 \\ R_2 & I_{n_2} \end{bmatrix} \cdot \begin{bmatrix} (I_{n_1} + R_1 R_2)^{-1} & 0 \\ 0 & (I_{n_2} + R_2 R_1)^{-1} \end{bmatrix}$$

is invertible. Therefore, $(A+E)-\lambda(B+F)$ is block diagonalizable by $P \cdot P_{EF}$ and $Q \cdot Q_{EF}$. This means that the spectra of the new diagonal blocks in (3.6) must be disjoint, since otherwise we could change E and F infinitesimally and destroy the block diagonalizability. This is true as long as $\|(E,F)\|_E$ satisfies the bounds in (3.12) and (3.13), implying that

$$\mathrm{diss}(\sigma_1, \sigma_2) \geq \frac{\min(\mathrm{Dif}_l(\sigma_1,\sigma_2), \mathrm{Dif}_u(\sigma_1,\sigma_2))}{(2(p^2+q^2))^{1/2} + 2(p \cdot q)^{1/2}} \geq \frac{\min(\mathrm{Dif}_l(\sigma_1,\sigma_2), \mathrm{Dif}_u(\sigma_1,\sigma_2))}{(2(p^2+q^2))^{1/2} + 2\max(p,q)}$$

This proves the first part of the theorem.

To compute a bound on p_{EF} we proceed as follows. First note that

$$\begin{bmatrix} I_n & Y \\ X & I_m \end{bmatrix}^{-1} = \begin{bmatrix} I_n & -Y \\ -X & I_m \end{bmatrix} \cdot \begin{bmatrix} (I_n - YX)^{-1} & 0 \\ 0 & (I_m - XY)^{-1} \end{bmatrix}$$

so that if $\|X\|_E < 1$ and $\|Y\|_E < 1$ we can estimate

$$\kappa\left(\begin{bmatrix} I_n & Y \\ X & I_m \end{bmatrix}\right) \leq \frac{(1+\max(\|X\|_E, \|Y\|_E))^2}{1-\|X\|_E \cdot \|Y\|_E} \quad . \tag{3.14}$$

Since

$$\max(\|L_1\|_E, \|L_2\|_E, \|R_1\|_E, \|R_2\|_E) \leq x < 1 \quad ,$$

we can apply (3.14) and Lemma 2 we get

$$p_{EF} \leq \kappa(PP_{EF}) \leq \kappa(P) \cdot \kappa(P_{EF}) \leq 2 \cdot p \cdot \frac{(1+x)^2}{1-x^2}$$

as claimed.

To get a bound on q_{EF} we repeat this entire argument using the following choices for P and Q:

$$P = \begin{bmatrix} I_{n_1} & L \\ 0 & I_{n_2} \end{bmatrix} \cdot \begin{bmatrix} q^{1/2} I_{n_1} & 0 \\ 0 & q^{-1/2} I_{n_2} \end{bmatrix}$$

and

$$Q = \begin{bmatrix} I_{n_1} & R \\ 0 & I_{n_2} \end{bmatrix} \cdot \begin{bmatrix} q^{1/2} I_{n_1} & 0 \\ 0 & q^{-1/2} I_{n_2} \end{bmatrix} \quad .$$

Q.E.D.

We record another result of this approach here, because we will need it in the next section. In it we use a measure of the distance between two subspaces $\mathbf{P}_1$ and $\mathbf{P}_2$ of equal dimension, the maximum angle $\theta_{max}(\mathbf{P}_1, \mathbf{P}_2)$, which we define in terms of the (acute) angle between nonzero vectors x and y ($\theta(x,y)$):

$$\theta_{max}(\mathbf{P}_1,\mathbf{P}_2) = \max_{\substack{x_i \in \mathbf{P}_i \\ x_i \neq 0}} \theta(x_1,x_2) \quad .$$

Lemma 7: Let $\|(E,F)\|_E$ satisfy the constraint in the statement of Theorem 3. Let $\mathbf{P}$ be the left deflating subspace of $A-\lambda B$ belonging to σ_1, and let $\mathbf{P}_{EF}$ be the corresponding left deflating subspace of $(A+E)-\lambda(B+F)$. Then

$$\theta_{max}(\mathbf{P},\mathbf{P}_{EF}) \le \arctan(x\cdot(p + \sqrt{p^2-1})) \quad .$$

If $x<1/2$ we can improve this bound to

$$\theta_{max}(\mathbf{P},\mathbf{P}_{EF}) \le \arctan(2\cdot x / p) \quad .$$

Similarly, if $\mathbf{Q}$ and $\mathbf{Q}_{EF}$ are right deflating subspaces of $A-\lambda B$ and $(A+E)-\lambda(B+F)$, respectively, corresponding to σ_1, then

$$\theta_{max}(\mathbf{Q},\mathbf{Q}_{EF}) \le \arctan(x\cdot(q + \sqrt{q^2-1})) \quad .$$

If $x<1/2$ we can improve this bound to

$$\theta_{max}(\mathbf{Q},\mathbf{Q}_{EF}) \le \arctan(2\cdot x / q) \quad .$$

Proof: We do only the case for the left deflating subspaces; the right one is analogous. The first n_1 columns in

$$P = \begin{bmatrix} I_{n_1} & L \\ 0 & I_{n_2} \end{bmatrix} \cdot \begin{bmatrix} p^{1/2} I_{n_1} & 0 \\ 0 & p^{-1/2} I_{n_2} \end{bmatrix}$$

span $\mathbf{P}$, or equivalently $[I_{n_1}|0]^T$ spans $\mathbf{P}$, and the first n_1 columns of PP_{EF} (P_{EF} as in the proof of Theorem 3), or equivalently

$$\begin{bmatrix} p^{1/2}\cdot I_{n_1} - p^{-1/2}\cdot L\cdot L_2 \\ -p^{-1/2}\cdot L_2 \end{bmatrix} \cdot [(I+L_1L_2)^{-1}]$$

spans $\mathbf{P}_{EF}$. Postmultiplying this set of columns by $(I+L_1L_2)\cdot(p^{1/2}\cdot I_{n_1} - p^{-1/2}\cdot L\cdot L_2)^{-1}$ does not change their span, and yields

$$\begin{bmatrix} I_{n_1} \\ -p^{-1}\cdot L_2\cdot(I_{n_1} - p^{-1}\cdot L\cdot L_2)^{-1} \end{bmatrix}$$

It is easy to show that the maximum angle between two spaces spanned by $[I|0]^T$ and $[I|Z]^T$ is $\arctan(\|Z\|)$ [Stewart1], so

$$\begin{aligned} \theta_{max}(\mathbf{P},\mathbf{P}_{EF}) &\le \arctan(\|-p^{-1}\cdot L_2\cdot(I_{n_1} - p^{-1}\cdot L\cdot L_2)^{-1}) \\ &\le \arctan(p^{-1}\cdot x / (1-p^{-1}(p^2-1)^{1/2})) \\ &\le \arctan(x\cdot(p + (p^2-1)^{1/2})) \quad . \end{aligned}$$

If $x<1/2$, $1-\|p^{-1}\cdot L\cdot L_2\|_E > 1/2$ and the result follows. Q.E.D.

To compare the strength of the results in Theorems 2 and 3 we need a lemma comparing Dif_u, Dif_l and Dif_λ:

Lemma 8:

$$\max\left(\mathrm{Dif}_u(\sigma_1,\sigma_2),\mathrm{Dif}_l(\sigma_1,\sigma_2)\right) \le \mathrm{Dif}_\lambda(\sigma_1,\sigma_2) \ .$$

Proof: We prove the result for Dif_u. The result for Dif_l follows from the symmetry between Dif_u and Dif_l.

$$\mathrm{Dif}_\lambda(A_{11},A_{22};B_{11},B_{22}) = \inf_{\substack{c,s \\ |c|^2+|s|^2=1}} \left(\sigma^2_{\min}(cA_{11}-sB_{11})+\sigma^2_{\min}(cA_{22}-sB_{22})\right)^{1/2}$$

$$= \inf_{\substack{c,s,u,v \\ |c|^2+|s|^2=1 \\ \|u\|=\|v\|=1}} \left\| \begin{bmatrix} cu^*A_{11}-su^*B_{11} \\ A_{22}cv-B_{22}sv \end{bmatrix} \right\|_E = \inf_{\substack{c,s,u,v \\ |c|^2+|s|^2=1 \\ \|u\|=\|v\|=1}} \left\| \begin{bmatrix} cvu^*A_{11}-svu^*B_{11} \\ A_{22}cvu^*-B_{22}svu^* \end{bmatrix} \right\|_E$$

$$= \inf_{\substack{c,s,P \\ |c|^2+|s|^2=1 \\ \mathrm{rank}(P)=1 \\ \|P\|=1}} \left\| \begin{bmatrix} cPA_{11}-sPB_{11} \\ A_{22}cP-B_{22}sP \end{bmatrix} \right\|_E \ge \inf_{\|(L,R)\|_E=1} \left\| \begin{bmatrix} LA_{11}-RB_{11} \\ A_{22}L-B_{22}R \end{bmatrix} \right\|_E$$

$$= \sigma_{\min} \begin{bmatrix} B_{11}^T\otimes I_{n_2} & A_{11}^T\otimes I_{n_2} \\ -I_{n_1}\otimes B_{22} & -I_{n_1}\otimes A_{22} \end{bmatrix} = \sigma_{\min} \begin{bmatrix} B_{11}\otimes I_{n_2} & -I_{n_1}\otimes B_{22}^T \\ A_{11}\otimes I_{n_2} & -I_{n_1}\otimes A_{22}^T \end{bmatrix} = \sigma_{\min} \begin{bmatrix} I_{n_2}\otimes B_{11} & -B_{22}^T\otimes I_{n_1} \\ I_{n_2}\otimes A_{11} & -A_{22}^T\otimes I_{n_1} \end{bmatrix}$$

(since $(X\otimes Y)^T=X^T\otimes Y^T$ and since there exists a permutation matrix P such that $P^T(X\otimes Y)P=Y\otimes X$)

$$= \sigma_{\min} \begin{bmatrix} I_{n_2}\otimes A_{11} & -A_{22}^T\otimes I_{n_1} \\ I_{n_2}\otimes B_{11} & -B_{22}^T\otimes I_{n_1} \end{bmatrix} = \mathrm{Dif}_u(A_{11},A_{22};B_{11},B_{22})$$

Q.E.D.

Applying this lemma and a little manipulation to the lower bounds on $\mathrm{diss}(\sigma_1, \sigma_2)$ in Theorems 2 and 3 shows that the lower bound in Theorem 2 is always stronger than the lower bound in Theorem 3:

Theorem 4:

$$\frac{\mathrm{Dif}_\lambda(\sigma_1,\sigma_2)}{\sqrt{2}\cdot(p+q)} \ge \frac{\min\left(\mathrm{Dif}_u(\sigma_1,\sigma_2)\ ,\ \mathrm{Dif}_l(\sigma_1,\sigma_2)\right)}{(2\,(p^2+q^2))^{1/2}+2\max(p,q)}$$

To see how much larger the one lower bound may be than the other, consider the example

$$A-\lambda B = \left[\begin{array}{cccc|cccc} \epsilon & 1 & & & & & & \\ & \cdot & \cdot & & & & & \\ & & \cdot & 1 & & & & \\ & & & \epsilon & & & & \\ \hline & & & & -\epsilon & 1 & & \\ & & & & & \cdot & \cdot & \\ & & & & & & \cdot & 1 \\ & & & & & & & -\epsilon \end{array}\right] - \lambda \cdot \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} - \lambda \begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix} ,$$

with both $A_{11}-\lambda B_{11}$ and $A_{22}-\lambda B_{22}$ being of dimension n. Then for small ϵ a little computation shows that the lower bound of Theorem 2 is proportional to ϵ^n and the lower bound of Theorem 3 is proportional to ϵ^{2n-1}, almost the square.

Even though Theorem 3 provides a worse lower bound on $\text{diss}(\sigma_1, \sigma_2)$, the analysis leading up to it allows us to bound the condition numbers of the P and Q matrices in (1.1), which the analysis of Theorem 2 does not allow.

Theorem 5: Let $\sigma = \bigcup_{l=1}^{n} \sigma_l$ be a disjoint partitioning. Let p_l and q_l denote the norms of the projectors onto the left and right deflating subspaces, respectively, belonging to σ_l. For $1 \le i \le b$ suppose

$$x_l \equiv \|(E,F)\|_E \cdot \frac{(2(p_l^2+q_l^2))^{1/2}+2\cdot\max(p_l,q_l)}{\min(\text{Dif}_u(\sigma_1,\sigma_2),\ \text{Dif}_l(\sigma_1,\sigma_2))} < 1$$

and let $x \equiv \max_l x_l$. Let P and Q denote matrices such that $P^{-1}((A+E)-\lambda(B+F))Q$ is block diagonal as in (1.1). Then P and Q can be chosen so that

$$\kappa(P) \le 2\cdot b \cdot \frac{(1+x)^2}{1-x^2} \cdot \max_{1\le l\le b} p_l$$

and

$$\kappa(Q) \le 2\cdot b \cdot \frac{(1+x)^2}{1-x^2} \cdot \max_{1\le l\le b} q_l$$

Proof: Let $p_l(EF)$ and $q_l(EF)$ denote norms of projectors onto left and right deflating subspaces belonging to σ_l, respectively, of $(A+E)-\lambda(B+F)$. By Lemma 2 and Theorem 3 we have

$$\begin{aligned} \kappa(P) &\le b \cdot \max_{1\le l\le b} p_l(EF) \\ &\le 2\cdot b \cdot \max_{1\le l\le b} \frac{(1+x_l)^2}{1-x_l^2} \cdot p_l \\ &\le 2\cdot b \cdot \frac{(1+x)^2}{1-x^2} \cdot \max_{1\le l\le b} p_l \ . \end{aligned}$$

An analogous sequence of inequalities show that

$$\kappa(Q) \le 2\cdot b \cdot \frac{(1+x)^2}{1-x^2} \cdot \max_{1\le l\le b} q_l \ .$$

Q.E.D.

Requiring that these bounds on $\kappa(P)$ and $\kappa(Q)$ be less than TOL yields the condition of Theorem A that the decomposition (1.1) be stable. This completes the proof of Theorem A.

Note that by Lemma 2 $b \cdot \max_i p_i$ is essentially the condition number of the best conditioned P and $b \cdot \max_i q_i$ the condition of the best conditioned Q that block diagonalize $A-\lambda B$. Therefore the extra factor $2(1+x)^2/(1-x^2)$ in Theorem 3 indicates how much the condition numbers of P and Q can grow beyond the minimum possible.

We note that if we specialize to the standard eigenproblem ($B=I$ and $\|A\| \le 1$), Theorem A yields essentially the same results as derived in [Demmel2] for the standard eigenproblem [Demmel and Kågström].

4. Singular Pencils.

Just as our analysis of regular pencils began with a triangular canonical form (Lemma 4), we also begin with such a canonical form for singular pencils:

Lemma 9: [Van Dooren3] given any pencil $A-\lambda B$ there exist unitary matrices P and Q such that $P^{-1}(A-\lambda B)Q$ is in the following quasi-uppertriangular form:

$$P^{-1}(A-\lambda B)Q = \begin{bmatrix} A_r-\lambda B_r & * & * \\ & A_{reg}-\lambda B_{reg} & * \\ & & A_l-\lambda B_l \end{bmatrix} \tag{4.1}$$

where $A_r-\lambda B_r$ has only L_l blocks in its KCF (i.e. all right minimal indices, hence the subscript r), $A_l-\lambda B_l$ has only L_j^T blocks in its KCF (i.e. all left minimal indices), $A_{reg}-\lambda B_{reg}$ is upper triangular and regular, and each * is an arbitrary conforming pencil. $A_{reg}-\lambda B_{reg}$ can be chosen with its spectrum on the diagonal in any order. Further, the blocks in the KCF of $A-\lambda B$ are precisely those which appear in the KCFs of $A_r-\lambda B_r$, $A_l-\lambda B_l$, and $A_{reg}-\lambda B_{reg}$.

Suppose now without loss of generality that our pencil is in the form of (4.1):

$$A-\lambda B = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} \lambda - \begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix} \tag{4.2}$$

where $A_{11}-\lambda B_{11}$ includes all of $A_r-\lambda B_r$ and possibly part of $A_{reg}-\lambda B_{reg}$, and $A_{22}-\lambda B_{22}$ contains all of $A_l-\lambda B_l$ and the remainder of $A_{reg}-\lambda B_{reg}$. In particular, we assume $A_{11}-\lambda B_{11}$ and $A_{22}-\lambda B_{22}$ contain disjoint parts (σ_1 and σ_2 respectively) of the spectrum σ of $A-\lambda B$. If we denote the numbers of rows and columns of $A_{ll}-\lambda B_{ll}$ by m_l and n_l, then it is easy to see $m_1 \le n_1$ (with equality if and only if $A_r-\lambda B_r$ is null) and $m_2 \ge n_2$ (with equality if and only if $A_l-\lambda B_l$ is null). We want to blockdiagonalize this pencil, the upper left block containing σ_1 and the lower right block σ_2. Equivalently, just as for the regular case, we seek P and Q such that

$$P^{-1}(A-\lambda B)Q = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix} - \lambda \begin{bmatrix} B_{11} & 0 \\ 0 & B_{22} \end{bmatrix}$$

with $A_{ll}-\lambda B_{ll}$ having the same KCF as $A_{ll}-\lambda B_{ll}$. The pair of reducing subspaces belonging to σ_1 are spanned by $Q_1=[I_{n_1}|0]^T$ and $P_1=[I_{m_1}|0]^T$ [Van Dooren3]. We seek to block diagonalize this pencil by choosing $P_2=[L^T|I_{n_2}]^T$ and $Q_2=[R^T|I_{n_2}]^T$, which leads to the equation

$$\begin{bmatrix} I_{m_1} & -L \\ 0 & I_{m_2} \end{bmatrix} \cdot \begin{bmatrix} A_{11}-\lambda B_{11} & A_{12}-\lambda B_{12} \\ 0 & A_{22}-\lambda B_{22} \end{bmatrix} \cdot \begin{bmatrix} I_{n_1} & R \\ 0 & I_{n_2} \end{bmatrix} = \begin{bmatrix} A_{11}-\lambda B_{11} & 0 \\ 0 & A_{22}-\lambda B_{22} \end{bmatrix}$$

or

$$A_{11}R - LA_{22} = -A_{12} \tag{4.3a}$$

$$B_{11}R - LB_{22} = -B_{12} \tag{4.3b}$$

which are identical to (3.2a) and (3.2b). Thus, we can rewrite (4.3ab) just as in the regular case as

$$\begin{bmatrix} I_{n_2}\otimes A_{11} & -A_{22}^T\otimes I_{m_1} \\ I_{n_2}\otimes B_{11} & -B_{22}^T\otimes I_{m_1} \end{bmatrix} \cdot \begin{bmatrix} \mathrm{col}R \\ \mathrm{col}L \end{bmatrix} \equiv Z_u \cdot \begin{bmatrix} \mathrm{col}R \\ \mathrm{col}L \end{bmatrix} = \begin{bmatrix} -\mathrm{col}A_{12} \\ -\mathrm{col}B_{12} \end{bmatrix} \tag{4.4}$$

This is a set of $2m_1n_2$ linear equations in $n_1n_2+m_1m_2$ unknowns, the entries of L and R. Since

$m_1 \le n_1$ and $m_2 \ge n_2$, we see we have at least as many unknowns as equations with equality if and only if $A-\lambda B$ is regular, the case analyzed in part 3 of this paper. When it is singular, it has a (nonunique) solution as stated in the following lemma:

Lemma 10: [Van Dooren3, lemma 2.3] (4.4) is a consistent set of equations for arbitrary A_{12} and B_{12} if:

1) $A_{11}-\lambda B_{11}$ has no left minimal indices in its KCF,
2) $A_{22}-\lambda B_{22}$ has no right minimal indices in its KCF, and
3) $\sigma_1=\sigma(A_{11}-\lambda B_{11})$ and $\sigma_2=\sigma(A_{22}-\lambda B_{22})$ are disjoint.

Thus, Z_u is of full rank leading us to define Dif_u just as for the regular case:

$$\mathrm{Dif}_u(A_{11},A_{22};B_{11},B_{22}) \equiv \sigma_{\min}(Z_u)$$

with the same trivial consequence:

$$\|(L,R)\|_E \le \|(A_{12},B_{12})\|_E \,/\, \mathrm{Dif}_u(A_{11},A_{22};B_{11},B_{22}) \ .$$

Also as before, Dif_u is specified merely by choosing σ_1 and $\sigma_2=\sigma-\sigma_1$, permitting us to write $\mathrm{Dif}_u(\sigma_1,\sigma_2)$ if $A-\lambda B$ is known from context or even just Dif_u if σ_1 is known as well.

Given a space of solutions of (3.3), we will choose the (L,R) of least norm since it leads to P and Q as well conditioned as possible. Call this minimum norm solution (L_0,R_0) and denote $(1+\|L_0\|^2)^{1/2}$ by p and $(1+\|R_0\|^2)^{1/2}$ by q. Just as Dif_u is specified only by σ_1, so are $\|L_0\|$, $\|R_0\|$, p and q.

We follow essentially the same approach as for Theorem 3 in section 3: Choose P and Q (using L_0 and R_0) to satisfy equation (3.1), leading to equation (3.5). Seek P_{EF}^{-1} and Q_{EF} of the same form as in section 3 such that premultiplying (3.5) by P_{EF}^{-1} and postmultiplying it by Q_{EF} blockdiagonalizes it. As before, this leads to equations (3.7ab) and (3.8ab) to solve. Again, we wish to apply Lemma 1. This time, however, the linear operators T in (3.9a) and (3.10a) are no longer square. To use Lemma 1, we need to show both operators have full rank. The linear operator T in (3.9a) is represented by Z_u in (4.4) above, which Lemma 10 showed to be full rank. The operator T of (3.10a) is represented using Kronecker products as

$$\begin{bmatrix} (A_{11}+E_{11})^T \otimes I_{m_2} & -I_{n_1} \otimes (A_{22}+E_{22}) \\ (B_{11}+F_{11})^T \otimes I_{m_2} & -I_{n_1} \otimes (B_{22}+F_{22}) \end{bmatrix} \ . \tag{4.5}$$

The next lemma shows that this matrix is of full rank.

Lemma 11: Assume $S_1-\lambda T_1$ has only L_k blocks and regular blocks in its KCF, that $S_2-\lambda T_2$ has only L_j^T and regular blocks in its KCF, and that σ_1 and σ_2 are disjoint, where σ_i is the spectrum of $S_i-\lambda T_i$. Then the system of linear equations

$$\begin{bmatrix} LS_1 - S_2R \\ LT_1 - T_2R \end{bmatrix} = \begin{bmatrix} C \\ D \end{bmatrix} \tag{4.6}$$

has at most one solution. Suppose $S_i-\lambda T_i$ is m_i by n_i. Then this unicity of solution is equivalent to the matrix

$$\begin{bmatrix} S_1^T \otimes I_{m_2} & -I_{n_1} \otimes S_2 \\ T_1^T \otimes I_{m_2} & -I_{n_1} \otimes T_2 \end{bmatrix} \tag{4.7}$$

having full rank.

Proof: The assumptions on the KCFs of $S_l-\lambda T_l$ imply that $m_1\le n_1$ and $m_2\ge n_2$. Thus, the matrix in (4.7) has at least as many rows as columns, so if we show (4.6) has at most one solution, this will imply (4.7) has full rank. Choose P_l and Q_l so that $P_l^{-1}(S_l-\lambda T_l)Q_l$ is in KCF. Call the blocks on the diagonal of the KCF $S_{lj}-\lambda T_{lj}$. Then (4.6) decomposes into set of independent equations

$$\begin{bmatrix} L'S_{1j}-S_{2k}R' \\ L'T_{1j}-T_{2k}R' \end{bmatrix} = \begin{bmatrix} C_{jk} \\ D_{jk} \end{bmatrix} . \tag{4.8}$$

If we show (4.8) has at most one solution for each j and k, we will be done. There are several cases. First suppose $S_{1j}-\lambda T_{1j}=L_{j}$. and $S_{2k}-\lambda T_{2k}=L_k^T$. Then it is easy to see from the forms of L_j. and L_k^T. that (4.8) is essentially triangular in that we first solve (4.8) for the first column of R', then the first column of L', then the second column of R', the second column of L', and so on. Thus, if (4.8) has a solution, it is uniquely determined. The second case is when $S_{1j}-\lambda T_{1j}$ is a Jordan block and $S_{2k}-\lambda T_{2k}=L_k^T$. In this case we may solve (4.8) successively for the last row of L', the last row of R', the next to last row of L' and so on. When $S_{1j}-\lambda T_{1j}$ is a block with an infinite eigenvalue, we return to the column by column regime. The other cases are similar, except when both $S_{1j}-\lambda T_{1j}$ and $S_{2k}-\lambda T_{2k}$ are regular. Since by assumption they have disjoint spectra, this reduces to the case covered in section 3. This proves that the matrix in (4.7) is of full rank under the conditions stated in the lemma. Q.E.D.

This lemma justifies the definition

$$\mathrm{Dif}_l(A_{11},A_{22};B_{11},B_{22}) \equiv \sigma_{\min}\left(\begin{bmatrix} A_{11}^T\otimes I_{m_2} & -I_{n_1}\otimes A_{22} \\ B_{11}^T\otimes I_{m_2} & -I_{n_1}\otimes B_{22} \end{bmatrix}\right)$$

As before we can show that this definition is really coordinate free allowing us to write $\mathrm{Dif}_l(\sigma_1,\sigma_2)$ when $A-\lambda B$ is known from context or just Dif_l if σ_1 is known as well. This leads us to the following extension of Lemma 7:

Theorem 6: Assume $A-\lambda B$ is in the form (4.2). Let $\mathbf{P}$ and $\mathbf{Q}$ be the left and right reducing subspaces of $A-\lambda B$ belonging to σ_1. Then if $(A+E)-\lambda(B+F)$ has reducing subspaces $\mathbf{P}_{EF}$ and $\mathbf{Q}_{EF}$ of the same dimensions as $\mathbf{P}$ and $\mathbf{Q}$, respectively, where

$$\|(E,F)\|_E = x\cdot\frac{\min\,(\mathrm{Dif}_u(\sigma_1,\sigma_2)\ ,\ \mathrm{Dif}_l(\sigma_1,\sigma_2))}{(2\cdot(p^2+q^2))^{1/2}+2\cdot\max\,(p,q)} \quad \text{where } x<1\ ,$$

then one of the following two cases must hold:

Case 1:

$$\theta_{\max}(\mathbf{P},\mathbf{P}_{EF}) \le \arctan(x\cdot(p+(p^2-1)^{1/2}))$$

and

$$\theta_{\max}(\mathbf{Q},\mathbf{Q}_{EF}) \le \arctan(x\cdot(q+(q^2-1)^{1/2}))\ .$$

If $x<1/2$ then we have the tighter bounds

$$\theta_{\max}(\mathbf{P},\mathbf{P}_{EF}) \le \arctan(2\cdot x\ /\ p)$$

and

$$\theta_{\max}(\mathbf{Q},\mathbf{Q}_{EF}) \le \arctan(2\cdot x\ /\ q)\ .$$

In other words, both angles are small, bounded above by a multiple of the norm of the perturbation $\|(E,F)\|_E$.

Case 2: Either

$$\theta_{max}(\mathbf{P},\mathbf{P}_{EF}) \geq \arctan\left(\frac{1}{\sqrt{m} \cdot p + (p^2-1)^{1/2}}\right)$$

or

$$\theta_{max}(\mathbf{Q},\mathbf{Q}_{EF}) \geq \arctan\left(\frac{1}{\sqrt{n} \cdot q + (q^2-1)^{1/2}}\right) .$$

In other words, at least one of the angles between perturbed and unperturbed reducing subspaces is bounded *away from* 0.

Proof: If $A-\lambda B$ were regular, the proof of case 1 would be Lemma 7. Since $A-\lambda B$ is of the form (4.2), $\mathbf{P}$ and $\mathbf{Q}$ are spanned by the columns of $[I_{m_1}|0]^T$ and $[I_{n_1}|0]^T$, respectively. In the course of proving Lemma 7 we assumed that $\mathbf{P}_{EF}$ was spanned by the columns of

$$\begin{bmatrix} I_{m_1} - p^{-1} \cdot L \cdot L_2 \\ -p^{-1} \cdot L_2 \end{bmatrix} . \tag{4.10}$$

First we will show that if this assumption is false, case 2 holds. If the space $\mathbf{P}_{EF}$ we wish to span does not contain a vector orthogonal to $\mathbf{P}$ (in which case case 2 trivially holds), then it is easy to see that $\mathbf{P}_{EF}$ can be spanned by the columns of $[I|X^T]^T$ for a suitable X. There are two cases: $I-XL$ is singular and $I-XL$ is nonsingular. If $I-XL$ is singular then some vector in $\mathbf{P}_{EF}$ is also in the space spanned by the columns of $[L^T|I]^T$. But if there were such a vector, then the maximum angle between $\mathbf{P}$ and $\mathbf{P}_{EF}$ would be at least $\arctan((p^2-1)^{-1/2})$, which implies that Case 2 holds. If $I-XL$ is nonsingular we claim that we can choose L_2 to make the columns of (4.10) have the same span as $[I|X^T]^T$: choose

$$L_2 = -p^{-1} \cdot (I-XL)^{-1} \cdot X$$

which, when substituted into (4.10) yields

$$\begin{bmatrix} I+L \cdot (I-XL)^{-1} \cdot X \\ (I-XL)^{-1} X \end{bmatrix} .$$

The top block of this last matrix is nonsingular, since

$$\begin{bmatrix} -X & I \\ I & 0 \end{bmatrix} \cdot \begin{bmatrix} L & I \\ I & 0 \end{bmatrix} = \begin{bmatrix} I & 0 \\ L \cdot (I-XL)^{-1} & I \end{bmatrix} \cdot \begin{bmatrix} I-XL & -X \\ 0 & I+L \cdot (I-XL)^{-1} \cdot X \end{bmatrix}$$

is nonsingular. Postmultiplying by the inverse of this block yields $[I|X^T]^T$.

Thus it suffices to consider potential new left reducing subspaces spanned by columns of the form (4.10) (and new right reducing subspaces spanned by analogous sets of vectors). This leads to equations (3.8ab) in the same way as in the proof of Theorem 3. The proofs of Theorem 3 and Lemma 7 apply here almost verbatim. The essential difference is that equations (3.7ab) and (3.8ab) are no longer square. Since equations (3.7ab) are underdetermined, we must use case 2 of Lemma 1 to estimate when they are soluble. From Lemma 1 we can get a solution of (3.7ab), but we lose uniqueness. Since equations (3.8ab) are overdetermined, we must use case 3 of Lemma 1 to estimate the solution when one exists; that one does exist is the assumption that reducing subspaces of the appropriate dimensions exist. The two cases correspond to the two cases in (2.3) and (2.4) of Lemma 1. Case 1 follows just as in Lemma 7, and corresponds to the bound on the solution in (2.3).

Case 2 corresponds to (2.4), which says that the solution of (3.8ab) satisfies

$$
\begin{aligned}
\|(L_2,R_2)\|_E &\geq \frac{1+(1-4\kappa)^{1/2}}{\|\phi\| \cdot \|T^+\|} \\
&\geq \frac{\mathrm{Dif}_l(\sigma_1,\sigma_2) - (2\cdot(p^2+q^2))^{1/2}\cdot\|(E,F)\|_E}{2\cdot q\cdot\|(E,F)\|_E} \\
&\geq \frac{\mathrm{Dif}_l(\sigma_1,\sigma_2) - (2\cdot(p^2+q^2))^{1/2}\cdot\dfrac{\mathrm{Dif}_l(\sigma_1,\sigma_2)}{(2\cdot(p^2+q^2))^{1/2}+2\cdot\max(p,q)}}{2\cdot q\cdot\dfrac{x\cdot\mathrm{Dif}_l(\sigma_1,\sigma_2)}{(2\cdot(p^2+q^2))^{1/2}+2\cdot\max(p,q)}} \\
&= \frac{\max(p,q)}{x\cdot q} \geq \frac{1}{x} \ .
\end{aligned}
$$

Thus, either $\|L_2\|_E \geq 2^{-1/2}/x$ or $\|R_2\|_E \geq 2^{-1/2}/x$. If $\|L_2\|_E \geq 2^{-1/2}/x$, then since $\mathbf{P}_{EF}$ is spanned by the columns of (4.10), one can see the lower bound on $\|L_2\|_E$ translates into the following lower bound on the tangent of the largest angle between $\mathbf{P}_{EF}$ and $\mathbf{P}$:

$$
\begin{aligned}
\tan(\theta_{\max}(\mathbf{P},\mathbf{P}_{EF})) &\geq \frac{p^{-1}\cdot\|L_2\|}{1+p^{-1}\cdot\|L\|\cdot\|L_2\|} \\
&= \left(\frac{p}{\|L_2\|}+(p^2-1)^{1/2}\right)^{-1} \geq \left(\sqrt{m}\cdot p+(p^2-1)^{1/2}\right)^{-1}
\end{aligned}
$$

as desired. If $\|R_2\|_E \geq 2^{-1/2}/x$, the proof is analogous. Q.E.D.

One way to interpret this lemma is as a nongeneric perturbation bound: if E and F are small enough, then for each pair of left and right reducing subspaces of $(A+E)-\lambda(B+F)$ of the same dimensions as the unperturbed pair, either both the left and right subspaces of $(A+E)-\lambda(B+F)$ will be a small angle (bounded by a multiple of the norm $\|(E,F)\|_E$) away from the unperturbed spaces, or else at least one of them will be bounded away in angle from the unperturbed space.

Theorem 6 also supplies perturbation bounds for algorithms used to compute the KCF. These algorithms compute decompositions of the form (4.1) or an equivalent form. The algorithms of [Van Dooren1], [Kågström1], and [Kågström2] among others are all stable in that they can produce an exactly singular pencil (along with its KCF) near the pencil $A-\lambda B$ supplied as input. The user may choose how near this exactly singular pencil must be to $A-\lambda B$ by varying certain thresholds in the algorithms. Of course if the pencil is square there may be no singular pencil within the distance chosen by the user, which the algorithm will then report (the algorithm may unfortunately fail to find such a matrix even if one exists). The algorithms are also stable in the sense that for suitably chosen thresholds and most input problems they will return pencils with the same singular structures in their KCFs for all input pencils sufficiently close to $A-\lambda B$. These properties are discussed at length in the references at the beginning of the paragraph.

Therefore, one can take the estimate of Theorem 6 and apply it to analyzing the error of standard algorithms in the following algorithm:

Algorithm 1:

1) Reduce a pencil $A-\lambda B$ to $A'-\lambda B'$ of the form (4.1) using a standard algorithm. Let δ be a bound for the perturbation the algorithm makes in the pencil in order to reduce it

(δ is computed by the algorithm). Let $\mathbf{P}'$ and $\mathbf{Q}'$ be left and right reducing subspaces of $A'-\lambda B'$ corresponding to σ_1.

2) Compute the bound

$$\Delta = \frac{\min(\mathrm{Dif}_u(\sigma_1,\sigma_2)\ ,\ \mathrm{Dif}_l(\sigma_1,\sigma_2))}{(2\cdot(p^2+q^2))^{1/2}+2\cdot\max(p,q)}$$

in the statement of Theorem 6.

3) If $\delta < \Delta$, then suppose $A''-\lambda B''$ is a singular pencil within distance $y < \Delta-\delta$ of $A-\lambda B$ with right and left reducing subspaces $\mathbf{P}''$ and $\mathbf{Q}''$ of the same dimensions as $\mathbf{P}'$ and $\mathbf{Q}'$, respectively. Then one of the two cases holds:

Case 1:

$$\theta_{max}(\mathbf{P}',\mathbf{P}'') \le \arctan(\frac{\delta+y}{\Delta}\cdot(p+(p^2-1)^{1/2}))$$

and

$$\theta_{max}(\mathbf{Q}',\mathbf{Q}'') \le \arctan(\frac{\delta+y}{\Delta}\cdot(q+(q^2-1)^{1/2}))\ ,$$

with better bounds if $A''-\lambda B''$ is within $(\Delta-\delta)/2$ of $A-\lambda B$.

Case 2: Either

$$\theta_{max}(\mathbf{P}',\mathbf{P}'') \ge \arctan(\frac{1}{\sqrt{m}\cdot p+(p^2-1)^{1/2}})$$

or

$$\theta_{max}(\mathbf{Q}',\mathbf{Q}'') \ge \arctan(\frac{1}{\sqrt{n}\cdot q+(q^2-1)^{1/2}})\ .$$

4) if $\delta > \Delta$, no perturbation bound can be made, because the perturbation δ made by the algorithm is outside the range Δ of our estimate.

Now let us interpret these results for the problem of computing controllability subspaces. Let $\mathbf{C}(C,D)$ denote the controllability subspace of the pair of matrices (C,D). As discussed in [Van Dooren2], this subspace is simply the left reducing subspace corresponding to $\sigma_1 = \varnothing$ of the pencil $A-\lambda B = [D|C-\lambda I]$. It is easy to see that this pencil can have neither infinite eigenvalues nor L_k^T blocks in its KCF since $B=[0|I]$ has full rank; hence it can only have finite eigenvalues and L_j blocks in its KCF. Also, one can see that the number of L_j blocks is a constant equal to the number of columns of D. Thus, the assumption in Theorem 6 about the perturbed pencil having reducing subspaces of the same size is implied by the assumption that the perturbed system $(C+E_C,D+E_D)$ has a controllability subspace of the same dimension as $\mathbf{C}(C,D)$. Also, the algorithms one used for this problem take advantage of the special form of $B=[0|I]$ by operating only on $A=[D|C]$ and so making no perturbation in B; thus we will assume $F=0$ [Van Dooren2]. Also, as we will see, R_2 and L_2 are quite simply related, allowing us to prove

Corollary 4: Suppose that

$$\dim(\mathbf{C}(C+E_C,D+E_D)) = \dim(\mathbf{C}(C,D))$$

and that

$$\|(E_C,E_D)\|_E = x\cdot\frac{\min(\mathrm{Dif}_u(\sigma_1,\sigma_2)\ ,\ \mathrm{Dif}_l(\sigma_1,\sigma_2))}{(2\cdot(p^2+q^2))^{1/2}+2\cdot\max(p,q)}\ ,\quad \text{where } x<1\ .$$

Then either

Case 1:

$$\theta_{max}(\mathbf{C}(C,D),\mathbf{C}(C+E_C,D+E_D)) \le \arctan(x\cdot(p+(p^2-1)^{1/2}))$$

with

$$\theta_{max}(\mathbf{C}(C,D),\mathbf{C}(C+E_C,D+E_D)) \le \arctan(2\cdot x\ /\ p)$$

if $x<1/2$, or

Case 2:

$$\theta_{max}(\mathbf{C}(C,D),\mathbf{C}(C+E_C,D+E_D)) \ge \arctan(\frac{1}{\sqrt{m}\cdot p+(p^2-1)^{1/2}}) \ .$$

Proof: It suffices to prove that $\|R_2\|_E = \|L_2\|_E$. This follows from (3.8b) and the special form of B and the fact that $F=0$. In fact, (3.8b) implies that $[0|L_2] = R_2$, from which the equality of their norms follows immediately. Q.E.D.

Algorithm 1 clearly also applies to algorithms for compute the controllability subspace. It is also easy to see that these results apply immediately to observability subspaces as well by duality [Wonham].

Another feature of a control system (C,D) for which we can derive perturbation bounds using this approach is the spectrum of the regular part, also called the *input decoupling zeroes* or *uncontrollable modes* of the control system.

Theorem 7: Suppose that Case 1 of Theorem 6 holds. Suppose further that the block $A_{22}-\lambda B_{22}$ is regular. (This implies $A-\lambda B$ has no L_k^T blocks in its KCF.) Then the spectrum of the perturbed pencil $(A+E)-\lambda(B+F)$ includes the spectrum of

$$(A_{22}+E'_{22})-\lambda(B_{22}+F'_{22})$$

where

$$\|(E'_{22},F'_{22})\|_E \le \sqrt{2}\cdot q\cdot \|(E,F)\|_E \ .$$

Similarly, if we instead assume $A_{11}-\lambda B_{11}$ is regular, then the spectrum of the perturbed pencil $(A+E)-\lambda(B+F)$ includes the spectrum of

$$(A_{11}+E'_{11})-\lambda(B_{11}+F'_{11})$$

where

$$\|(E'_{11},F'_{11})\|_E \le \sqrt{2}\cdot p\cdot \|(E,F)\|_E \ .$$

Proof: From (3.6) we see that

$$E'_{11} = E_{11}-L_1E_{21} \ ,$$

$$E'_{22} = E_{22}-L_2E_{12} \ ,$$

$$F'_{11} = F_{11}-L_1F_{21} \ , \text{and}$$

$$F'_{22} = F_{22}-L_2F_{12} \ .$$

Substitute in the bounds on $\|E_{ij}\|$ and $\|F_{ij}\|$ found there along with $\|L_i\| \le 1$ in these equations. Q.E.D.

We may now use estimates from section 3 of this report, [Stewart2], [Stewart3] or anywhere else to bound the perturbations in the spectrum of the pencil. In the case of

uncontrollable modes of a control system this theorem implies

Corollary 5: Assume we are in Case 1 of Corollary 4. Then the uncontrollable modes of the perturbed control system $(C+E_C, D+E_D)$ are the eigenvalues of the pencil

$$(A_{22}+E'_{22})-\lambda B_{22}$$

where

$$\| E'_{22} \|_E \leq \sqrt{2} \cdot q \cdot \| (E_C, E_D) \|_E \ .$$

In [Van Dooren2], P and Q are chosen in (4.1) so that $P^{-1}BQ = [I|0]$, implying $B_{22}=I$. Thus the problem of finding the eigenvalues of the perturbed pencil $(A_{22}+E'_{22})-\lambda B_{22}$ above reduces to perturbation theory for the standard eigenproblem.

5. Numerical Examples.

Example 1: Consider the regular pencil of the introduction for $\eta = 10^{-5}$:

$$A - \lambda B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 10^{-5} \end{bmatrix} - \lambda \cdot \begin{bmatrix} 10^{-5} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} .$$

First we consider the partitioning of the spectrum σ of this pencil into $\sigma_1 = \{10^5\}$, $\sigma_2 = \{1\}$, and $\sigma_3 = \{10^{-5}\}$. We wish to know for what values of ϵ and TOL this decomposition is stable. From the definitions of Dif_u and Dif_l in section 3, we see that

$$\mathrm{Dif}_u(\sigma_1, \sigma - \sigma_1) = .998 = \mathrm{Dif}_l(\sigma_1, \sigma - \sigma_1) \ ,$$

$$\mathrm{Dif}_u(\sigma_2, \sigma - \sigma_2) = 5 \cdot 10^{-6} = \mathrm{Dif}_l(\sigma_2, \sigma - \sigma_2) \ , \text{ and}$$

$$\mathrm{Dif}_u(\sigma_3, \sigma - \sigma_3) = 5 \cdot 10^{-6} = \mathrm{Dif}_l(\sigma_3, \sigma - \sigma_3) \ .$$

It is also easy to see that $p_i = q_i = 1$ for $i = 1,2,3$. Therefore, from (1.3) and (1.4) of Theorem A we see that if

$$\epsilon < \frac{5 \cdot 10^{-6}}{4} = 1.25 \cdot 10^{-6}$$

and

$$2 \cdot 3 \cdot \frac{(1 + \epsilon \cdot 8 \cdot 10^5)^2}{1 - (\epsilon \cdot 8 \cdot 10^5)^2} < TOL$$

then the decomposition $\sigma = \sigma_1 \bigcup \sigma_2 \bigcup \sigma_3$ will be stable. For example, if we choose $TOL = 100$, then ϵ has to satisfy

$$\epsilon < 1.1 \cdot 10^{-6}$$

for stability.

If we set $TOL = \infty$, then since

$$\mathrm{Dif}_\lambda(\sigma_1, \sigma - \sigma_1) = .99999$$

$$\mathrm{Dif}_\lambda(\sigma_2, \sigma - \sigma_2) = 7.07 \cdot 10^{-6}$$

$$\mathrm{Dif}_\lambda(\sigma_3, \sigma - \sigma_3) = 7.07 \cdot 10^{-6}$$

we see from (1.5) that the decomposition $\sigma = \sigma_1 \bigcup \sigma_2 \bigcup \sigma_3$ will be stable if

$$\epsilon < 2.5 \cdot 10^{-6} \ .$$

From Corollary 2, we see that these upper bounds on ϵ could not possibly be any larger than $7.07 \cdot 10^{-6}$, since a perturbation of that size would make the eigenvalues 0 and 10^{-5} coalesce.

If $\epsilon > 2.5 \cdot 10^{-6}$, then Theorem A no longer guarantees the the eigenvalues at 0 and 10^{-5} cannot coalesce. In this case, we consider the decomposition $\sigma_1 = \{10^5\}$ and $\sigma_2 = \{0, 10^{-5}\}$. Now the quantities we need to apply Theorem A are

$$\mathrm{Dif}_u(\sigma_1, \sigma_2) = \mathrm{Dif}_u(\sigma_2, \sigma_1) = \mathrm{Dif}_l(\sigma_1, \sigma_2) = \mathrm{Dif}_l(\sigma_2, \sigma_1) = .998 \ ,$$

$p_i = q_i = 1$, and

$$\mathrm{Dif}_\lambda(\sigma_1, \sigma_2) = .99999 \ .$$

From (1.3) and (1.4) we see $\sigma = \sigma_1 \bigcup \sigma_2$ is stable if

$$\epsilon < .2495$$

and

$$2 \cdot 2 \cdot \frac{(1+4.008 \cdot \epsilon)^2}{1-(4.008 \cdot \epsilon)^2} < TOL \quad .$$

For example, if $TOL = 100$, then the decomposition is stable if

$$\epsilon < .2303 \quad .$$

If we set $TOL = \infty$, then by (1.5) $\sigma = \sigma_1 \bigcup \sigma_2$ is stable if

$$\epsilon < .3535 \quad .$$

From Corollary 2 we see that these upper bounds on epsilon can be no larger than .99999, since a perturbation of that size would make the eigenvalues 10^5 and 10^{-5} coalesce. If $\epsilon > .3535$ Theorem A cannot guarantee that σ can be decomposed at all.

Example 2: We consider the singular pencil

$$A - \lambda B = [D | C - \lambda I] = \begin{bmatrix} 10^{-5} & 1-\lambda & 0 & 0 \\ 0 & 0 & 2-\lambda & 0 \\ 0 & 0 & 0 & 3-\lambda \end{bmatrix} .$$

This pencil is in the canonical form (4.2) with $m_1 = 1$, $m_2 = 2$, $n_1 = 2$ and $n_2 = 2$. The left reducing subspace **P** which is spanned by $[1,0,0]^T$ is the controllability subspace $\mathbf{C}(C,D)$ of the system (C,D).

The quantities of interest in Corollary 4 are

$$\mathrm{Dif}_u(\varnothing,\{2,3\}) = \mathrm{Dif}_l(\varnothing,\{2,3\}) = .112 \quad ,$$

and $p = q = 1$. Thus Corollary 4 implies that if E_C and E_D are such that

$$1 = \dim(\mathbf{C}(C,D)) = \dim(\mathbf{C}(C+E_C, D+E_D))$$

and

$$\| (E_C, E_D) \|_E = .028 \cdot x \qquad \text{where } x < 1$$

then either

Case 1:

$$\theta_{max}(\mathbf{C}(C,D) \ , \ \mathbf{C}(C+E_C, D+E_D)) \le \arctan(x)$$

or

Case 2:

$$\theta_{max}(\mathbf{C}(C,D) \ , \ \mathbf{C}(C+E_C, D+E_D)) \ge \arctan(1/\sqrt{3}) = \arctan(.577) = .52 \quad .$$

Thus, for example, if we have $\| E_C, E_D) \|_E < .001$, then any one dimensional controllability subspace of $(C+E_C, D+E_D)$ will either be within .036 radians of $\mathbf{C}(C,D)$ or at least .52 radians away from $\mathbf{C}(C,D)$. A simple example of the latter situation is

$$[D+E_D | C+E_C - \lambda I] = \begin{bmatrix} 0 & 1-\lambda & 0 & 0 \\ 10^{-5} & 0 & 2-\lambda & 0 \\ 0 & 0 & 0 & 3-\lambda \end{bmatrix}$$

in which case the new controllability subspace is spanned by $[0,1,0]^T$ and so is orthogonal to $\mathbf{C}(C,D)$.

In case the perturbed controllability subspace falls into Case 1, then we can use Corollary 5 to bound the uncontrollable modes of the perturbed system: the perturbed modes

are eigenvalues of the matrix

$$\begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} + E'$$

where $\| E' \|_E \leq \sqrt{2} \cdot \| (E_C, E_D) \|_E$. Gershgorin's theorem supplies the simple bound that the perturbed uncontrollable modes lie in disks centered at 2 and 3 with radii $\sqrt{2} \cdot \| (E_C, E_D) \|_E \leq x \cdot .04$.

6. References.

[Demmel1] J. Demmel, "The Condition Number of Equivalence Transformations that Block Diagonalize Matrix Pencils," SIAM J. Num. Anal., Vol. 20 (1983) pp. 599-610

[Demmel2] J. Demmel, "Computing Stable Eigendecompositions of Matrices," submitted to Linear Alg. App.

[Gantmacher] F. Gantmacher, *The Theory of Matrices*, Vol. II, (Transl.) Chelsea, New York, (1959)

[Demmel and Kågström] J. Demmel and B. Kågström, "Stable Eigendecompositions of Matrix Pencils $A-\lambda B$," Report UMINF-118.84, Institute of Information Processing, Umeå University, Umeå, Sweden, June, 1984

[Kågström1] B. Kågström, "The Genererized Singular Value Decomposition and the General $(A-\lambda B)$ Problem," BIT 24 (1984) 568-583

[Kågström2] B. Kågström, "RGSVD - An Algorithm for Computing the Kronecker Structure and Reducing Subspaces of Singular $A-\lambda B$ Pencils," Report UMINF-112.83, Institute of Information Processing, Umeå University, Umeå, Sweden (1983)

[Moler and Stewart] C. Moler and G. Stewart, "An Algorithm for the generalized eigenvalue problem," SIAM J. Num. Anal., (1973) pp. 241-256

[Rosenbrock] H. Rosenbrock, *State Space and Multivariable Theory*, Nelson, London, 1970

[Stewart1] G. Stewart, "Error and Perturbation Bounds for Subspaces Associated with Certain Eigenvalue Problems," SIAM Review, Vol. 15, (1973) pp. 752-764

[Stewart2] G. Stewart, "Gershgorin Theory for the Generalized Eigenproblem $Ax=\lambda Bx$," Math. of Comp., Vol. 29, No. 130, April 1975, pp. 600-606

[Stewart3] G. Stewart, "On the Sensitivity of the Eigenvalue Probem $Ax=\lambda Bx$," SIAM J. Num. Anal., Vol. 9, No. 4, Dec. 1972, pp. 669-686

[Van Dooren1] P. Van Dooren, "The Computation of Kronecker's Canonical Form of a Singular Pencil," Lin. Alg. Appl., Vol. 27, (1979), pp. 103-141

[Van Dooren2] P. Van Dooren, "The Generalized Eigenstructure Problem in Linear System Theory," IEEE Trans. Auto. Control, Vol. AC-26, No. 1, February 1981, pp. 111-128

[Van Dooren3] P. Van Dooren, "Reducing Subspaces: Definitions, Properties and Algorithms," in Matrix Pencils, Proceedings, Piteå Havsbad, 1982 Lecture Notes in Mathematics Vol. 973, Eds. A. Ruhe and B. Kågström, Springer Verlag, Berlin, 1983

[Van Dooren4] P. Van Dooren, "ALGORITHM 590: DSUBSP and EXCHQZ: Fortran Routines for Computing Deflating Subspaces with Specified Spectrum," ACM TOMS, Vol. 8, No. 4, (1982) pp. 376-382

[Varah] J. Varah, "On the separation of two matrices," SIAM J. Num. Anal., Vol. 16, No. 2, (1979) pp 216-222

[Waterhouse] W. Waterhouse, "The Codimension of Singular Matrix Pairs," Lin. Alg. Appl., 57:227-245 (1984)

[Wilkinson] J. Wilkinson, "The Differential System $B\dot{x}=Ax$ and the Generalized Eigenvalue Problem $Au=\lambda Bu$," Nat. Phys. Lab. Report NAC 73, Jan. 1977

[Wonham] M. Wonham, *Linear Multivariable Theory. A Geometric Approach*, 2nd ed., Springer Verlag, New York, 1979

NYU COMPSCI TR-163 c.2
Demmel, James Weldon

Computing stable
eigendecompositions